2008

Rational-empirical strategy with IRT to derive personality disorder scales from the personality assessment inventory / by Leah Burgess.

Burgess, Leah

A Rational-Empirical Strategy with IRT to derive Personality Disorder Scales from the Personality Assessment Inventory

Leah Burgess

Ph.D. Dissertation

Department of Psychology

Lakehead University

2008

Your file Votre référence
ISBN: 978-0-494-47148-7
Our file Notre référence
ISBN: 978-0-494-47148-7

Table of Contents

II. Structural

III. Present Study

List of Appendixes

Abstract

The Personality Assessment Inventory (PAI) is a broadband measure of adult psychopathology that assesses two of ten personality disorders (PDs). A rational-empirical strategy using expert raters and Item Response Theory based analyses (Samejima's Graded Response Model) was employed in an attempt to construct an additional eight PD scales from the existing PAI item pool. Raters demonstrated strong agreement in identifying PAI items that capture all 10 PDs. The IRT analyses supported that the PAI items can be reconfigured to assess the 10 PDs, and convergence with nonparametric scalability coefficients and internal consistency reliability estimates was demonstrated. Also, preliminary discriminant and convergent validity evidence per correlations with Millon's (1997) PD scales was generally consistent with extant PD research. Overall, the results demonstrated several advantages of applying IRT methods to psychopathology research.

A Rational-Empirical Strategy with IRT to derive Personality Disorder Scales from the Personality Assessment Inventory

The proposed study is a test development exercise intended to enhance the clinical utility of an existing broad-band measure of adult personality and psychopathology: the Personality Assessment Inventory (PAI). In its present form, the PAI only directly assesses two discrete personality disorders (PDs), however, current nosologies of mental disease contain classifications for ten different PDs. Moreover, research has consistently indicated that individuals with a personality disorder tend to display characteristics or traits consistent with more than one disorder (Costa & Widiger, 2002a). To facilitate accurate diagnosis and treatment planning, there is an existing need in clinical settings to comprehensively screen patients for several different PDs. Although, alternative narrow measures exist (e.g., Schedule for Nonadaptive and Adaptive Personality (SNAP); Clark, 1993a), practitioners unequivocally prefer (or at least utilize) broadband measures (e.g., Camara, Nathan, & Puente, 2000). The PAI is gaining widespread clinical appeal because its psychometric properties are arguably superior to alternative tools (cf. Butcher et al., 2001; Morey, 1991). The rationale for the proposed study is that the PAI may not be getting used to its full potential. Through applying a rational-empirical strategy to derive additional PD scales from the PAI, it is proposed that the measurement precision and clinical utility of the existing test can be improved. If successful, traditional administration of the PAI will yield substantially more clinical information that can then be used to readily inform more targeted assessment, diagnosis, and treatment planning in a time- and cost-effective manner.

As frequently cited and adopted by others (e.g., Jackson, 1970), one of the most scientifically rigorous approaches to scaling is to mirror as closely as possible Loevinger's (1957) tri-stage model of test development. The quality of any test is most simply estimated by examining evidence of reliability and validity. However, reliability and validity evidence assumes several forms, and there are various subcomponents of psychometric evidence that are requisite precursors to demonstrating satisfactory evidence of reliability and validity. The strength of Loevinger's model is that it conceptualizes test development across all stages of the evidence gathering process. The three stages she identifies are termed substantive, structural, and external, which Loevinger boldly describes as "mutually exclusive, exhaustive,...and mandatory" (p. 653-654).

The substantive stage necessitates that the exercise be grounded in a well-articulated theoretical framework and secondly, that test items be constructed such that they reflect a representative sample of the broad universe of relevant content. For two reasons, the proposed study cannot uphold this ideal. First, although the proposed test development exercise is grounded in theory, the veridicality of the theory can be challenged on several grounds. Second, the proposed test items will be culled from an existing battery and, in turn, may not adequately sample the respective content/theoretical domains. Although Loevinger's ideal strategy cannot wholly be matched, it will be approximated whenever possible. The rationale for the modified strategy here is short-term practical utility. As previously stated, the PAI is a viable alternative to psychometrically weaker broadband measures of psychopathology, but its assessment of PD is too limited in scope. Indeed, lack of consideration of a greater number of PDs has been described as a "substantial disadvantage of the PAI" (p. 413, Widiger & Coker, 2002).

In terms of theory, the Diagnostic and Statistical Manual of Mental Disorders - Text Revision (*DSM-IV-TR*; American Psychiatric Association, 2000) is the adopted theoretical framework for the proposed scale development exercise. In North America, the *DSM* system is the most widely accepted nomenclature and nosology of mental disorders for both research and clinical purposes. Selecting the *DSM* framework was a difficult decision. With respect to theoretical orientation, personality psychology and the domain of personality disorders in particular is a very dissentious field. Members are strongly divided along theoretical as well as scholar versus practitioner lines. Given that scale development for latent constructs is an inherently difficult process under optimal circumstances, it is especially difficult to formulate a theoretical foundation for the proposed scale development exercise against the backdrop of this divisive field. Thus, although the *DSM* model of PDs is adopted as the theoretical foundation of the proposed scales, there is no concomitant assumption that the *DSM* model is optimal or that construct validity evidence for the *DSM* has yet been sufficiently demonstrated. Rather, the *DSM* model is adopted on the basis of consistency with current standards of applied clinical practice and because a viable alternative does not yet exist.

Loevinger's (1957) second or structural stage of test development refers to the requisite process of employing empirical, quantitative methods to evaluate the fidelity of the observed scale structure against the intended model. Structural investigations are primarily concerned with item level analyses. For the proposed scales, IRT methods will be adopted for this purpose. Measures of normative and disordered personality functioning have traditionally been derived from scale construction and statistical analyses grounded in classical test theory (CTT) methodology. However, alternative method theories exist that have the potential to generate more accurate assessment tools. In particular, as reviewed by Embretson and Reise (2000), IRT

is one such method that has demonstrated marked empirical precision and practical utility. Given the potential wealth of psychometric information that can be gleaned through IRT compared to CTT methods, and the evidence to suggest that particular forms of IRT analyses have the potential to generate scales with superior external validity without employment of a stratified random sampling design (Lord & Novick, 1968; Wright, 1967), IRT based methods will be used to empirically assess the internal structure of the proposed PD scales.

Loevinger's (1957) third or external stage of test development is akin to the traditional notion of external validity or the generalizability of the results, as well as the traditional notion of construct validity or the amassing of evidence to support that the test does indeed assess the construct it was intended to measure. IRT methods will again be used to begin the external validation process. It is beyond the scope of this project to adequately address this domain. Independent validation efforts over time will be required to generate sufficient breadth and depth of empirically derived confirmatory or refutatory evidence. As an initial step in this process, once the respective PD scales have been generated, IRT modeling will again be applied to assess discriminant and convergent evidence for construct validity. IRT statistical modeling programs allow item responses from various scales to be compared. In this way, responses to the newly derived PD scales will be compared against existing measures, such as the PD scales of the Minnesota Multiphasic Personality Inventory – second edition (MMPI-2; Morey, Waugh, Blashfield, 1985; Somwaru & Ben-Porath, 1995). A more detailed outline of the proposed test development project and explanation of the rationale for each of the various stages follows.

SUBSTANTIVE

Personality Disorders

Definition

The construct of personality disorder has proven elusive to define. Several forms of definitions have been postulated over time. Original definitions were largely derived from clinical insights, but have since evolved to incorporate more empirically determined concepts. As reviewed and cited by Livesley (2001) and Jablensky (2002), terms including "manie sans delire[1]" (Pinel, 1809), moral insanity (Maudsley, 1874; Pritchard, 1835), psychopathic personalities (Koch, 1891; Schneider, 1923), "formes frustes[2]" (Kraepelin, 1907), and character armor (Reich, 1933) are the dominant precursors of current terminology. In addition, Millon (1981) identifies and cites the following precursors: biological/physiological based theories (Hirt, 1902; Kretschmer, 1925; Sheldon, 1940) and theories of pathological temperament (e.g., Kahn, 1928; Sjobring, 1914; Tramer, 1931). Linehan (1993) adds Stern's (1938) conceptualization of borderline neuroses. Each of these theories has since evolved into the more familiar theoretical orientations which encompass the most recent models of personality disorder: psychoanalytic, interpersonal, personological, multivariate, and empirical orientations (Wiggins, 2003; Winter & Barenbaum, 1999).

Unfortunately, despite the accumulating years of clinical theorizing and empirical investigation, no explicit operational definition of personality or personality disorder exists either within or across the respective theoretical orientations. Livesley (2001), however, suggests that despite the lack of a satisfactory operational definition, consensus across orientations does exist

---

[1] Manie sans delire = mania without delusions (Jablensky, 2002, p. 113).
[2] Formes frustes = a less severe form of a nonpersonality based mental illness (Livesley, 2001, p. 9).

as to the two primary elements of a definition: Personality reflects (a) regularity or "consistency in thinking, perceiving, feeling", and behaving and (b) the "configuration," "integration", and "organization" of the affiliated human traits and attributes (p. 7). Thus, personality appears to reflect a dynamic, but constrained construct. The notion of personality disorder builds on this foundation and, in a related way, is concerned with identifying (a) consistency and "regularities in personality pathology" and (b) failures of organization and integration (Livesley, 2001, p. 7). The primary obstacle in articulating an operational definition of PD is clarifying the distinction between normal and pathological personality functioning.

Normal versus Pathological

There is no criterion in any realm (e.g., biological, psychological, or sociological) that consistently and conclusively differentiates normal from pathological. Nonetheless, some form of differentiation needs to be established in order to inform diagnosis, treatment, and further research. In the PD literature, several strategies have been proposed (Livesley, 2001; Millon, 1981; Strack & Lorr, 1994). However, each strategy can likely be subsumed by the superordinate structure used to define all other forms of psychopathology. In the broadest sense, as reviewed by Davison and Neale (1996), pathology or abnormality can be defined through "statistical infrequency, violation of norms, personal distress, disability or dysfunction, and unexpectedness" (p. 6). As the authors caution, no single method is entirely satisfactory, and determining a criterion within each method is also somewhat arbitrary and changes over time.

Livesley (2001) provides a cogent summary of the normal-abnormal debate as applied specifically to definitions of personality pathology. Livesley identifies five strategies employed by PD researchers/clinicians to identify personality pathology: One, Axis I attenuation models:

PDs are a less extreme variant of Axis I pathology. Two, normal personality intensification models: PDs are a more extreme variant of normal personality traits. Three, social deviance and learning models: PD is the result of inappropriate socialization and/or "adaptive failure" (p. 9). Four, deficit models: PDs comprise structural or functional impairments. Five, unique personality orientation models: PDs are unique configurations of normal and abnormal traits or characteristics. Elements of Davison and Neale's (1996) definitions of pathology are readily evident in the application of each of these schemes. However, with the exception of statistical infrequency, none of the overarching PD schemes explicitly demarcates the boundary between normal and abnormal. Further, the schemes are not mutually exclusive because some are simply descriptive while others attempt to incorporate causal or etiological processes. More recently, perspectives on how best to conceptualize normal versus pathological personality functioning appear most centrally focused on the debate over dimensional versus categorical PD models. A more explicit discussion of the categorical-dimensional debate, and the more prominent theories or models of PD follows.

Models of PD

Dimensional vs. Categorical

Strack and Lorr (1994) distill the schemes of personality pathology into frameworks highly similar to Livesley's (2001), but they explicitly emphasize the perspective of dimensional versus categorical conceptualization. They identify four schemes. The first encompasses the purest form of the categorical perspective. Normal and pathological personality are conceived as discrete, "categorically distinct entities" owing to discrepant pathogenesis that can be identified through objective criteria (p. xvi). The second encompasses the purest form of the dimensional

perspective. All forms of personality functioning are conceived as dimensional. Normal and pathological personality are variants of the same traits. Personality pathology is identified as the arbitrary threshold on a continuum of overall personality functioning that is demarcated by the elusive criterions of subjective distress, inability to function, or statistically above or below average. The third proposes that systematic differences in and combinations of dimensional traits yield discrete, categorically distinct personality disorders. Finally, the fourth holds that normal personality is dimensional, but pathological personality is "categorically distinct". PDs result from the interaction between biological vulnerabilities and normative, dimensional traits (p. xvi). The third and fourth schemes reflect amalgamations of both categorical and dimensional frameworks. Although dissidents remain (e.g., Haslam, 2003), results of recent and sophisticated quantitative research has fairly overwhelming supported dimensional schemes (e.g., O'Connor & Dyce, 1998; O'Connor, 2002a, 2002b, 2005a). Moreover, even Axis I domains of psychopathology as assessed on popular measures appear to share the same dimensional structure across clinical and nonclinical populations (O'Connor, 2002a). Ongoing disagreements over proposed categorical and dimensional models appear to be highly fueled by misunderstandings of or differences in interpretation of the definition of categorical and dimensional.

More specifically, the available categorical and dimensional schemes are not necessarily mutually exclusive (Blashfield, 1993; Clark, Livesley, & Morey, 1997; Kraemer, Noda, & O'Hara, 2003; Skinner, 1986), particularly when understood from the perspective of more recent conceptualizations of taxonomy (Meehl, 2004). For example, once diagnosis is required "categorical and dimensional approaches are fundamentally equivalent" (Kraemer et al., p.18). If any dichotomous indicators of category membership assume the form of ordered rankings of at

least three indicators, the scheme can be considered dimensional (Kraemer et al.). As well, categorical taxonomies or valid, largely discrete taxa can conceivably be explained by a single dimensional indicator (e.g., trait), several dimensional indicators, or unique combinations of shared dimensional indicators (Meehl, 2004). Consequently, as emphasized by O'Connor (2002), it is an oversimplification to conceive of all nosological systems with discrete PD diagnoses as indicative of purely categorical taxonomies or alternative models derived from trait theory of basic personality functioning that do not explicitly include diagnostic classifications of PD as exclusively dimensional.

Although empirical research has convincingly rejected purely categorical models of PD (Clark, Watson, & Reynolds, 1995; O'Connor, 2002a; O'Connor & Dyce, 1998) it has not, consequently, yielded an operational definition of PD or identified an alternative dimensional or hybrid model(s) that is clearly superior. As will be explained later in greater detail, the common variance of proposed PD dimensional frameworks is well accounted for by a dimensional model of basic personality functioning or traits (O'Connor, 2005a). However, a model of basic personality functioning is considered by many PD researchers/clinicians to be inadequate for clinical purposes, in particular, diagnosis and treatment planning (e.g., Benjamin, 1993; Clark, 1993). Consequently, it is the specification of the remaining qualities of proposed dimensional models of PD, those believed to encompass qualities outside the domain of basic personality functioning/traits, that lack consensus and continue to be debated. Thus, it appears that the core of a dimensional model for PD is now reasonably well established and is consistent with dimensional models of basic personality functioning. Flushing out the remaining qualities of a PD model so that it is sufficiently comprehensive with respect to being readily applicable for clinical and research purposes remains outstanding.

Based on clinical utility and empirical findings, the two most prominent categorical and dimensional PD schemes are the *DSM* and the Five-Factor Model (FFM), respectively. Each is described in greater detail below, followed by a brief overview of competing models.

DSM Model

In North America, the most widely accepted nomenclature and nosology of mental disorders for both research and clinical purposes is the *DSM-IV-TR* (2000). The current conceptualization of PDs is actually a fairly recent phenomenon. Evidence of the current scheme was not noticeable until publication of the *DSM-III* in 1980 and by publication of the *DSM-III-R* in 1987 was virtually identical. The *DSM-IV-TR* categorizes disorders on a multiaxial system which was intended to delineate all "domains of information" that conceivably contribute to accurate diagnosis and treatment, including environmental stressors and medical conditions. The multiaxial system was also intended to promote widespread "application of the biopsychosocial model" of psychopathology (*DSM-IV-TR*, p. 27). All major forms of clinical psychopathology are categorized on Axis I, but the PDs are classified separately on Axis II[3]. Owing to their unique status distinct from Axis I, PDs are often conceived as a relatively homogeneous, discrepant form of psychopathology with unique pathogenesis. However, this perception is not consistent with either the intent of the *DSM-IV-TR* classification system or current etiological models of PD (e.g., Coccaro, 2001; Depue, & Lenzenweger, 2001).

Because Theodore Millon served on the Task Force for development of the *DSM-III*, which demarcated the first inclusion of the multiaxial system, his opinion regarding the intent of categorizing PDs on an Axis separate from more traditional forms of psychopathology (Axis I)

[3] Note that the *DSM-IV-TR* Axis II also encompasses Mental Retardation spectrum diagnoses. In this document, however, any reference to Axis II is only intended to reference personality disorders.

warrants consideration. Millon (1981) states that aside from delineating significant personality dysfunction as a form of mental disorder in its own right, the inclusion of personality disorders on a discrete axis was actually intended to force practitioners to both consider and conceptualize traditional forms of psychopathology (Axis I disorders) within the larger context of an individual's global personality functioning. Millon's perspective remains consistent with the *DSM*'s description of the current multiaxial system. It is also consistent with current empirical evidence that suggests PDs are reflective of rigid and extreme variations of normal personality traits as opposed to some form of unique personality dysfunction (O'Connor, 2005b; O'Connor & Dyce, 2001).

Adhering to *DSM-IV-TR* (2000) classifications, the three guideposts comprising objective indexes of pathology across virtually all disorders are (a) internal dysfunction, (b) pain, suffering, or distress, and (c) impaired functioning or failure to adapt to mainstream society. The *DSM-IV-TR* defines PD as a chronic, deviant, rigid, omnipresent means of thinking, feeling, interacting, and behaving that "leads to distress or impairment" with an onset by early adulthood (p. 685). Thus, *DSM-IV-TR* appears to conceptualize PD as either a dysfunction at the level of one's personality or as simply a dysfunctional or maladaptive personality in and of itself. Also, the *DSM-IV-TR* acknowledges that dysfunction or psychopathology occurs when normative personality traits become overly rigid and maladaptive for a given environment. Hence, the *DSM* endorses (but does not operationalize) a trait theory perspective of PDs (O'Connor & Dyce, 2001). Furthermore, the *DSM-IV-TR* delineates a two-step process in diagnosing PDs which, as Livesley (2001) suggests, highlights elements of both the common clinical presentation across all PDs and the substantial heterogeneity under the PD rubric.

The *DSM* is also a hierarchical system where 10 discrete PDs are grouped under three superordinate clusters: "odd or eccentric"; "dramatic, emotional, or erratic"; and "anxious or fearful", which are labeled Clusters A, B, and C, respectively (APA, 2000, p. 685-686). Within this system, to diagnose PD individuals must first satisfy the "general diagnostic criteria" (APA, p. 689). The general criteria includes (a) presence of a rigid personality pattern evidenced in at least two specific domains: cognition, affect, interpersonal functioning, and/or impulse control; (b) early onset, enduring/chronic, and (c) pervasive dysfunction or distress. If met, individuals must secondly satisfy diagnostic criteria for a specific PD (e.g., schizoid, histrionic). This takes the form of dichotomous decision making on the presence or absence of a threshold listing of symptoms (e.g., 5 of 9 symptoms must be present). A brief description of each PD grouped by its respective cluster is outlined in Table 1.

Table 1

Description of DSM Personality Disorders by Hierarchical Cluster

Cluster[a]	Disorder (Abbreviation)	•Symptom Description
A	Paranoid (PAR)	-"pervasive distrust and suspiciousness of others"...[others'] motives are interpreted as malevolent" (p. 276)[b]
A	Schizoid (SZD)	-"pervasive pattern of detachment from social relationships and a restricted range of expression of emotions in interpersonal settings" (p. 277)
A	Schizotypal (SZT)	-"pervasive pattern of social and interpersonal deficits marked by acute discomfort with, and reduced capacity for, close relationships...and cognitive or perceptual distortions and eccentricities of behavior" (p. 279)
B	Antisocial (ANT)	-"pervasive pattern of disregard for and violation of the rights of others occurring since age 15 years" (p. 279)
B	Borderline (BOR)	-"pervasive pattern of instability of interpersonal relationships, self-image, and affects, and marked impulsivity" (p. 280)
B	Histrionic (HIS)	-"excessive emotionality and attention seeking" (p. 281)
B	Narcissistic (NAR)	-"grandiosity (in fantasy or behavior), need for admiration, and lack of empathy" (p. 282)
C	Avoidant (AVD)	-"social inhibition, feelings of inadequacy, and hypersensitivity to negative evaluation" (p. 283)
C	Dependent (DEP)	-"excessive need to be taken care of that leads to submissive and clinging behavior and fears of separation" (p. 284)
C	Obsessive-Compulsive (COM)	-"preoccupation with orderliness, perfectionism, and mental and interpersonal control, at the expense of flexibility, openness, and efficiency" (p. 285)
n/a	Personality Disorder Not Otherwise Specified (PDNOS)	-"for disorders of personality functioning that do not meet criteria for any specific [PD]" (p. 286) -Examples include: mixed PD symptom presentation and associated functional impairment, but the PD symptoms are insufficient to reach threshold within any single PD criteria set; satisfy a PD diagnosis not included in the *DSM*

Note. [a]Cluster A = "Odd or eccentric"; B = "Dramatic, emotional, or erratic"; C = "Anxious or fearful" (APA, 2000, p. 685-686); [b]All references are to APA (1994).

Criticisms of the DSM model. Upon initial review of this model, the *DSM* system appears comprehensive and sufficient. However, the model is routinely criticized on several accounts. The major points of controversy follow. Broadly, the criticisms of the *DSM* can be subsumed by the overarching concern of insufficient evidence of construct validity. One of the core criticisms repeatedly raised is that the original Axis II conceptualizations were derived from clinical or expert opinion with limited augmentation from empirical research. As well, the scientific basis of the on-going decision making process for selecting disorders to include (or remove) is weak (Westen, 1997). Although this is slowly changing (e.g., Blais & Norman, 1997), the criticism still largely holds because none of the *DSM-IV-TR* PD categories are derived from the results of rigorous, rational-empirical research strategies (Widiger & Trull, 1998). For example, the existing literature was reviewed and field trials were conducted by work groups specifically convened to empirically inform construction of the *DSM-IV* PD criteria. However, the actual field trials were conducted on the existing *DSM-III-R* criteria sets as opposed to the proposed revisions for the *DSM-IV*. Extant research on the *DSM-III/R*[4] and results of the field trials were simply used to inform the decision-making process in creating the *DSM-IV* PD categories/diagnoses and criteria sets. Alternatively stated, the criteria sets now included in the *DSM-IV-TR* were not subjected to a field trial (Widiger & Trull).

Also, results of independent convergent and discriminant validity investigations repeatedly yield mixed findings (e.g., Blackburn, Donnelly, Logan, & Renwick, 2004; Blais, Benedict, & Norman, 1998; Clark, Livesley, & Morey, 1997; Grilo et al., 2001; Grilo & McGlashan, 2000). Two of the most comprehensive large scale studies (Morey, 1988; Blais et al.) that examined all 11 PDs in a given study, yielded similar conclusions. Using clinical samples of 291 and 320, respectively, the authors found that the DSM-III-R criteria sets

[4] DSM-III/R = *DSM-III* and *DSM-III-R*.

demonstrated satisfactory convergent validity evidence, mixed but fair internal consistency estimates (median $\alpha = .68$) and poor discriminant validity evidence. As well, as reviewed by Widiger and Costa (1994) and Wiggins and Pincus (2002), the structure of the hierarchical system or the three cluster dimensions have not been well supported across factor analytic investigations. Lastly, findings are inconsistent across different measures and methods (Clark, Livesley, & Morey; Trull, 1993; Dyce, O'Connor, Parkins, & Janzen, 1997). Overall, as emphasized by Wiggins and Pincus (2002), the inconsistent findings across studies of construct validity are commonly attributed to the lack of consensus on a gold standard index of personality pathology to serve as the comparison criterion across validity investigations.

Excessive comorbidity is another frequent criticism. Reviews by Bornstein (1998), Clark, Watson, and Reynolds (1995), Costa and Widiger (2002), Livesley (2001), and Widiger and Coker (2002), highlight the severity of this problem. Comorbidity rates across PDs, are in the range of 67-85%, regardless of the measure or type of assessment used. Individuals who meet criteria for one PD typically satisfy criteria for at least three or four more. And, PD-NOS is the most frequently used diagnostic category. True comorbidity should not exceed chance (McCrae, 1994). Further, a review by Pfol (1999) indicates that comorbidity rates are also excessively high between Axis I and II disorders. Rates in clinical populations are approximately 50% and nonclinical populations range from 10-15%. A separate review by Dolan-Sewell, Krueger, and Shea (2001) suggests that individuals with PD have a 66-97% chance of having a comorbid Axis I condition, whereas 13-81% of individuals with Axis I may have a comorbid PD.

Widiger and Coker (2002) also note a compelling observation: Evolutions in the *DSM* Axis I criteria sets have seemingly contributed to increased comorbidity rates. As a specific example, they describe how social phobia used to be consistent with a circumscribed, specific

anxiety disorder. However, *DSM-IV* incorporated a "generalized subtype" with early onset, pervasive "history of social inhibition or shyness", and a lifelong symptom duration (APA, 2000, p. 453). Consequently, Widiger and Coker conclude that "there is no longer any meaningful distinction between a social phobia and avoidant personality disorder" (p. 423). Others have also highlighted the shared symptom overlap among BOR and bipolar disorder criteria sets. Thus, in some instances, it appears that excessive comorbidity rates are indeed simply diagnostic or statistical artifacts. Moreover, longitudinal investigations have repeatedly documented that certain normative and maladaptive personality traits precede the development of various Axis I conditions (Dolan-Sewell et al., 2001). Thus, the boundary between Axis I and II is inherently fuzzy.

As emphasized by Jablensky (2002), it is inconceivable that the observed comorbidity rates between Axis I and II and, particularly, within Axis II reflect true diagnostic comorbidity which is defined as the "simultaneous presence of two (or more) *aetiologically independent* conditions [italics original] (p. 114). As Lilienfeld et al. (1994) state, the comorbid term is frequently misapplied which erroneously "encourages the premature reification of diagnostic entities" (p. 71). More likely, the high comorbidity rates reflect the commonly expressed concerns of overly inclusive diagnostic schemes, lack of theoretical underpinning(s), and polythetic criteria sets. Drawing from the definition of superordinate categories espoused by cognitive psychologists (e.g., Rehder, 2006), the PD rubric should, at minimum, serve to constrain the number of permutations of PD features that are permissible across all of the individual disorders. As highlighted by others (e.g, Hurt et al., 1990; Widiger, 1998), given the *DSM* diagnostic format (e.g., any 5 of 9 symptoms), the heterogeneity of clinical presentation within any PD diagnosis can be substantial. Moreover, although the *DSM-IV-TR* prefaces that the

PD criteria sets are presented in order of "decreasing diagnostic importance" (p. 686), all indicators are given equal weight. Hence, any intended sense of prototypical presentation is lost in applied practice.

On the other hand, attempts to raise the internal consistency of the diagnoses have, unfortunately, yielded categories that are possibly overly exclusive or redundant: Criteria for some PDs have been criticized for being too narrow in scope. For example, Westen and Shedler (1999) describe how the criteria for Paranoid PD "are essentially seven indices of a single trait, chronic mistrust" (p. 274). Several additional and relevant facets of personality pathology are inherently lacking when a diagnosis is virtually reduced to pathology on a single trait. Similarly, a strong tradition of empirical research on ANT (and psychopathy) consistently reveals that one hallmark symptom of ANT is lack of empathy (e.g., Harper, Hakstian, & Hare, 1998; Hare, 2006; Hare, Kiehl, et al., 2004; Soderstrom, 2003). However, this and other more personality trait oriented indicators were reportedly dropped from inclusion in the *DSM-IV* because they were believed to be more difficult to objectively assess and might potentially increase comorbidity with NAR (Hare & Hart, 1995). However, the obvious cost of applying this strategy to increase reliability is a potential loss in construct validity – a seemingly unsound decision. In sum, it appears that the *DSM* PD criteria sets have not yet attained an appropriate balance of depth and breadth in scope.

The last set of criticisms encompass the issue of less than optimal clinical utility. The *DSM-IV* categories have been criticized for being poor predictors of response to treatment (Livesley, 2001) and personality traits rather than PD diagnoses can be stronger predictors of treatment seeking and functional impairment or severity of distress (Clarkin, Hull, Cantor, & Sanderson, 1993; Westen and Arkowitz-Westen, 1998). There are direct clinical consequences

for failing to maximally operationalize trait concepts in the criteria sets. Individuals who have elevations in PD domains but are not diagnostic, cannot be described/diagnosed via Axis II even though personality symptoms may be central to their dysfunction and/or assessment and treatment planning (Westen, 1997). Lack of *DSM* diagnosis has financial and occupational implications because alternative diagnostic schemes are not necessarily recognized by clinician regulatory bodies or insurers. Also, the dichotomous, present/absent symptom indicator scheme has been criticized for creating a loss in the richness of a given clinical presentation and ignoring strengths (Jablensky, 2002). As well, it has been argued that additional domains of functioning believed to be associated with pathological personality are lacking in the *DSM* conceptualizations. Livesley has repeatedly stated that a core component of PD not reflected in the *DSM* scheme is a reference to impairment in the functional capacity to integrate and organize all facets of one's personality. As well, *DSM* PD concepts have been described as too disparaging and, hence, possibly too difficult for clients to endorse or accept (Schacht, 1993). Although many of these final criticisms are not based on empirical investigations, because I endorse the perspective that clinical application is the ultimate goal of any research program related to PD assessment, it is likely worthwhile to at least acknowledge and take into consideration the concerns raised by practicing clinicians.

Strengths of the DSM model. Despite seemingly pervasive concerns for the validity of the *DSM* PD scheme, it is not without merit. Sophisticated taxometric investigations have yielded support for some of the existing *DSM* PD criteria sets, most notably SZD and ANT (Haslam, 2003). Several of the *DSM* PD criteria sets demonstrate differential temporal stability. This lends support to the *DSM* PD diagnostic scheme because the findings are not entirely

accounted for by regression to the mean (or health) effects (Morey et al., 2004). Also, unique combinations of FFM facet scores have demonstrated to differentially predict various *DSM* PD diagnoses (O'Connor & Dyce, 2002), which provides evidence to refute claims that the *DSM* PD categories are entirely redundant. As well, reviews of treatment outcome studies for Axis I pathology indicate that comorbid *DSM* PD diagnoses are differentially related to treatment outcome (Steketee, Chambless, & Tran, 2001). Note as well that a more detailed discussion of related reliability and validity evidence is later presented in the *assessment of PD* section. Lastly, Pfhol (1999) reminds critics to reconsider several unwarranted assumptions they may hold about Axis II. Specifically, Pfhol states that a PD diagnosis does not necessarily "exclude syndromes that show genetic or familial relationships to Axis I disorders, lessen in severity after several decades, respond to medications, or relate to abnormalities in neurotransmitter systems that may also be relevant to Axis I syndromes" (p. 89). In sum, regardless of the surrounding controversy and less than optimal construct validity evidence, the *DSM* PD scheme remains the diagnostic standard for practitioners in North America, and its application has direct and immediate implications for access to treatment and financial compensation.

Five Factor Model

The current exemplification of the FFM is the product of a lengthy, cumulative, and rigorous multi-investigator research history that originated in the 1930s and progressed sporadically through the 1960s until finally gaining widespread recognition and acceptance in the 1980s (Digman, 1990, 2002; Wiggins & Pincus, 2002). It is built on a foundation of factor analytic investigations of trait adjectives. The pioneering works of McDougall (1932) and Allport and Odbert (1936) are typically credited as the original contributions to FFM theory. As

reviewed in Digman's and Piedmont's (1998) historical accounts of the origins of the FFM, it is one of the most robust findings in personality psychology. Eliciting dimensions consistent with the FFM is repeatedly found across different measures, raters, ages, and methods. Yet, despite being postulated over 70 years ago, owing to the popularity and dominance of grand theory paradigms in personality research throughout history, it was not until the computer era (which permitted user-friendly application of factor analytic investigations) that trait models were widely tested. Subsequently, the convergence on five factors was relatively consistently replicated across studies. Hence, the FFM model was only widely accepted after 1980.

Based on historical reviews of the published literature on trait theory (Digman, 1990), no single author is credited with creating the FFM in and of themselves. As cited by Digman, credit appears to be dispersed across several researchers who made notable contributions (e.g., Borgatta, 1964; Cattell, 1957; Eysenck, 1970; Fiske, 1949; Goldberg, 1981; Norman, 1963; Tupes & Christal, 1961). At present, the most prominent representation of the FFM is Costa and McCrae's (1992a) depiction that is operationalized in their Revised NEO Personality Inventory (NEO-PI-R). The five domains of the FFM are termed: extraversion, agreeableness, conscientiousness, neuroticism, and openness. Although there is consistency across the literature with respect to the general personality construct area that is subsumed by each domain, there is less consensus on the most appropriate descriptive term for each domain. Most notably, the conscientiousness and openness labels are debated (Digman, 1990). In addition, the FFM model is hierarchical, so each primary domain subsumes several lower order facets. A summary of the facets grouped by respective domain is illustrated in Table 2. Although evidence for construct validity is accumulating, there is less consensus over the nature of the facets in comparison to the broad domains (Costa & Widiger, 1994, 2002b).

Table 2

Domains and lower order facets of the FFM

Domain	Facet
Neuroticism	Anxiety
	Anger/Hostility
	Depression
	Self-Consciousness
	Impulsiveness
	Vulnerability
Extraversion	Warmth
	Gregariousness
	Assertiveness
	Activity
	Excitement seeking
	Positive emotions
Openness	Fantasy
	Aesthetics
	Feelings
	Actions
	Ideas
	Values
Agreeableness	Trust
	Straightforwardness
	Altruism
	Compliance
	Modesty
	Tendermindedness
Conscientiousness	Competence
	Order
	Dutifulness
	Achievement striving
	Self-discipline
	Deliberation

Note. Costa & Widiger (1994)

The FFM is proffered from the trait theory orientation in personality psychology. Owing to a rich empirical research tradition, substantial evidence for the construct validity of the trait concept has amassed (Funder, 2001). However, of interest here is the applicability of the trait concept as operationalized in the FFM to abnormal personality functioning. As several reviewers have highlighted, once personality psychology gained credence as a unique discipline, abnormal and personality psychologists traditionally worked relatively independently. It is only within the last 10 to 20 years that concerted research efforts that integrate both domains have been undertaken. As it turns out, in this short period of time trait theory has indeed markedly informed our understanding of PDs. Building on the pioneering initiatives of Wiggins and Pincus (1989), Costa and McCrae (1990), Borkenau and Ostendorf (1990), and Widiger, Frances, Harris, Jacobsberg, Fyer, and Manning (1991); O'Connor and Dyce's recent work (O'Connor & Dyce, 1998; Dyce & O'Connor, 1998; O'Connor & Dyce, 2001; O'Connor, 2002a, 2002b, O'Connor & Dyce, 2002; O'Connor, 2005a, 2005b) has resolved, to the extent empirical methods of latent psychological constructs permit, several key issues central to the debate around the applicability of the FFM to disordered personality functioning.

More specifically, largely as a result of dissatisfaction with the evidence for construct validity of the *DSM* framework, several researchers have debated the applicability of various competing models of normal and abnormal personality functioning for the spectrum of PD. As emphasized by O'Connor (2002a), results from independent attempts to identify the true structure of PDs have yielded conflicting findings. O'Connor and Dyce (1998) were the first to actually statistically test the applicability of the various models on the same data sets that spanned clinical and nonclinical populations and different assessment measures. Since that original work, several follow-up studies have revealed additional insights. Key findings include

the following: First, normal and pathological personality traits appear to exist within the same dimensional/structural universe. It is unlikely that PDs are qualitatively distinct phenomena (O'Connor, 2002a, 2005a). Second, of the available models of normal and abnormal personality, the FFM of normal personality provides the best-fitting model of the underlying structure of PDs (O'Connor & Dyce, 1998, 2002). Third, the difference between normal and abnormal personality is likely a matter of rigidity and extremity or severity, rather than kind (O'Connor & Dyce, 2001). Note, however, that more recent analyses suggest that the traits of PDs may not be as extreme as originally conceived (O'Connor, 2005b). Finally, the FFM is a broad bandwidth model derived through data reduction techniques that are intended to simplify complex phenomena through explaining shared variance. Overall, the FFM appears well supported as a core dimensional framework underlying normative and disordered personality functioning.

Criticisms of the FFM. The two key criticisms of the FFM are that it is too general or broad in scope with respect to basic personality functioning and, alternatively, too narrow in scope with respect to personality disorder functioning: It fails to comprehensively provide the necessary and sufficient information needed for clinical assessment and intervention (Benjamin, 1993; Butcher & Rouse, 1996; Kernberg, 1996; Tellegen, 1993). More specific criticisms levied at the theoretical level include (a) excessive reliance on laypersons' conceptualization of personality traits and statistical properties to establish validity (Schacht, 1993) and (b) failure to adequately incorporate emotion or affective characteristics (Tellegen, 1993). The harshest critics (Butcher & Rouse, 1996) equate the FFM with folk wisdom, decry it is unscientific, and characterize its proponents as procrustean. Similar to concerns raised with the *DSM* model, the clinically utility of the FFM has also been questioned. Schacht suggests that the FFM is simply

too descriptive, lacks depth, and lacks an explanation of when and how extreme trait scores are causally linked to clinically significant "distress and dysfunction" (p. 116). Investigators have also identified a host of other personality variables deemed relevant for clinical situations that are not adequately addressed by the FFM. These reflect more theoretical orientation specific domains such as, individuation and negative valence (Ben-Porath & Waller, 1992), cognitive processes (sense of self, object relations), intrapsychic structures and processes or "morphologic organization" (p. 145), and biophysical domains (mood, temperament) (Millon & Davis, 1996; Kernberg, 1996). The respective orientation specific criticisms are not reviewed here because the list is as long as the depth of personality theory across the various orientations.

Ultimately, as discussed by O'Connor (2002b), because the FFM is a broad bandwidth model derived through data reduction techniques that are intended to simplify complex phenomena through explaining shared variance, it is well supported as a dimensional model underlying PD diagnoses. However, this concomitantly implies that the model is likely too inclusive to diagnose specific PDs or to sufficiently discriminate among PDs for clinical purposes. This is not a weakness of the model: The purpose of the FFM is to identify and account for broad dimensions as opposed to "scores on specific measures" or clinical diagnoses (O'Connor, p. 1999). The facets of the FFM have, however, demonstrated to "substantially increase specificity and discrimination between PDs" (O'Connor & Dyce, 2002, p. 243). Regardless, as O'Connor (2002b) explains, even with additional study facet level predictions will not be able to account for more variance than that explained by the primary factor solution of any given measure. Thus, the FFM is a viable core model for the trait parameter in PDs. However, it appears to remain insufficient as a diagnostic scheme for PDs for applied clinical purposes.

Competing Models of Personality Disorder

Several additional theorists have proposed different means or criteria sets to aid differentiating normal from abnormal personality functioning. As reviewed by O'Connor and Dyce (1998), models of PD have been derived through theoretical speculation, clustering by related symptoms, and by identifying relations with nonclinical personality traits and correlations with "personality test scores" (p. 3). A comparison of the more prominent, contemporary models of PD are outlined in Table 3. The original intent of this summary table was to delineate models of PD as opposed to models of basic personality. However, upon reviewing the literature, it becomes clear that many theorists and classification systems do not create unique models for normal personality versus pathological. Finally, this is not intended to be an exhaustive list of all PD models. Some historically noteworthy, grand personality theories are not reviewed because they do not lend themselves to empirical testing and are not readily used in current PD assessment practice (e.g., Freudian and Jungian psychodynamic theories; personological, existential, and humanistic theories).

Table 3

Description of the Prevalent Models of PD

Author and Model	Description
Cattell •16 Factor Model (Cattell & Cattell, 1995; Conn & Rieke, 1994)	-trait theory -personality consists of 16 primary or surface traits that are subsumed by five global or second order source traits (extraversion, anxiety, tough-mindedness, independence, self-control) -the 16 primary traits are as follows: warmth, reasoning, emotional stability, dominance, liveliness, rule-consciousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehension, openness to change, self-reliance, perfectionism, tension
Cloninger (1987) •Three-Factor Model	-biosocial theory -PDs are rooted in a biogenetic deficit and can be attributed to maladaptive temperament reactions -personality/PD is most aptly captured by three core temperament domains: novelty seeking, harm avoidance, and reward dependence
Cloninger and Svrakic (1994) •Seven-Factor Model	-expansion of Cloninger's three-factor model and derived from evolving factor analytic evidence -PD is most aptly captured by four genetically determined temperament domains (novelty seeking, harm avoidance, reward dependence, and persistence) and three self-concept determined character domains (self-directedness, cooperativeness, and self-transcendence)
Eysenck (1947, 1952) •Two-Factor Model •Three-Factor Model	-trait theory -original model defined personality by two trait domains: neuroticism and extroversion -later added a third trait domain, psychoticism
Kernberg (1996) •Psychoanalytic Model	-ego psychology and object relations theory -PDs can be mapped onto three core domains: psychotic, borderline, and neurotic personality organizations

Table 3 continued

Description of the Prevalent Models of PD

Model and Author	Description
Kiesler (1986, 1996), Leary (1957), Wiggins (1985) •Interpersonal Circumplex	-interpersonal theory -personality is most aptly characterized by a series of orthogonal, interpersonal traits captured by two primary axes: "control (dominance vs. submissiveness) and affiliation (friendliness vs. hostility)" (Kiesler, 2004, p. 1)
Millon (1969, 1981) •Circumplex PD Model	-Biosocial Learning Theory -Evolutionary Theory -personality is conceived on three dimensions of reinforcement that reflect instrumental behaviors (active-passive), motivations (pleasure-pain), and systemic or source influences (self-other) -the reinforcement domains interact with four coping styles: dependent, independent, ambivalent, and detached to yield 11 PDs
Tellegen •Three-Factor Model •Four-Factor Model (Patrick, Curtin, & Tellegen, 2002; Tellegen, 2006)	-psychobiological trait theory -personality is most aptly captured by three orthogonal domains consisting of two temperament factors (negative and positive emotionality) and one behavioral inhibition factor (constraint) -later added an additional primary trait domain (absorption)

Summary of PD Models

On the basis of either widespread clinical application or empirical findings, the *DSM* and the FFM are the most popular and viable diagnostic schemes. In terms of direct clinical application, however, as reviewed, neither is without significant flaws. This is not unexpected. Livesley (2001) ultimately proposes that no definition or model of PD will prove convincing until it satisfies the following criteria: specifies and explicates (a) "the defining features", (b) the dissimilar qualities "from other mental disorders", (c) the "derivation from normal personality",

and (d) the dissimilarity between disordered and normal personality" (pp. 9-10). Many of Livesley's arguments, although warranted, may be unrealistic or unattainable at this time. His specifications of the necessary and sufficient components for a conclusive definition of PD seem unrealistic because many components ultimately confront the same obstacle that underlies all study of any form of psychopathology: A definition of mental disorder. As of yet, a conclusive or satisfactory operational definition of mental disorder does not exist.

Second, although trait theorists have near conclusively delineated normal personality functioning, boundary zones demarcating pathology for each respective trait have not been identified (O'Connor, 2005b). And, even more perplexing, clear distinctions between personality and other psychological constructs (e.g., mood, affect) also have yet to be conclusively determined (Akiskal, 1994; Maremmani et al., 2005). Consequently, it seems unrealistic to expect PD researchers and clinicians to be able to delineate clear boundary and inclusionary criteria for PDs at this point in time. Finally, it has been logically demonstrated that no form of psychopathology can satisfy the three conditions necessary for a valid taxometric analysis, namely: (a) "indicators must be valid representations of the disorder, (b) covariation between indicators for reasons other than the construct of interest must be minimal, and (c) each indicator used in a given taxometric analysis should represent a phenotypically distinct dimension of symptom of the disorder" (Cole, 2004, p. 5). Our extant knowledge of PDs and psychopathology is simply not yet sufficient to adequately satisfy these criteria. Therefore, as typical across all domains of applied psychology, researchers and clinicians investigating PD must proceed while withholding the assumption that their elective approach is entirely valid. More specifically here, because immediate clinical utility is the primary outcome goal, the *DSM* model is adopted as the

grounding framework while fully acknowledging that this strategy has both theoretical and practical limitations.

Personality Disorder Assessment

Overview

Pre *DSM-III*, projective tests dominated personality assessment, and following the *DSM-III/R* a host of more objective assessment tools including, semi-structured interviews and broad- and narrow-band self-report rating scales, began to proliferate[5]. Surveys of clinicians' assessment practices (Camara et al., 2000; Holaday, Smith, & Sherry, 2000; Piotrowski, 1999; Piotrowski, Sherry, & Keller, 1985; Rabin, Barr, & Burton, 2005; Watkins, Campbell, Nieberding, & Hallmark, 1995) indicate that next to a clinical interview, clinicians appear to favor the use of broadband measures of psychopathology that integrate PD assessment with other forms of Axis I psychopathology, as opposed to measures explicitly designed to assess personality traits or PDs. Secondly, on average, clinicians appear to somewhat favor tradition over sound psychometric properties in selection of their tools. In particular, since at least 1971 (Lubin, Wallis, & Paine), the Rorschach and the Minnesota Multiphasic Personality Inventory (second edition; MMPI-2) continue to be the most widely used measures of psychopathology. Furthermore, other projective techniques also remain highly popular (Thematic Apperception Test (TAT), sentence completion tests, drawing tests).

It is noteworthy that the widespread usage of projectives and the dominance of the MMPI-2 continue to persist despite evidence that suggests these measures have questionable validity across several different applications (e.g., Costa, Zonderman, McCrae, & Williams,

[5] Note that although PDs can be diagnosed in individuals under 18 years of age and age-appropriate psychopathology measures exist for children and adolescents, the focus here is on assessment tools for adult populations.

1992; Helmes & Reddon, 1993; Johnson, Butcher, Null, & Johnson, 1984; Simms, Casillas, Clark, Watson, & Doebbeling, 2005; Waller, 1999; Wood, Garb, Lilienfeld, & Nezworski, 2002). Based on their review of extant survey data, Camara and colleagues (2000) conclude that since the 1960s clinicians appear to predominantly utilize a core assessment battery that encompasses a Wechsler intelligence test, the MMPI/2, and the Rorschach or Thematic Apperception Test. Indeed, surveys of practitioners in the 1990s and 2000s indicate that clinicians (and neuropsychologists) continue to rank the MMPI/2 and a Wechsler test as their first or second most commonly administered tool (e.g., Watkins et al., 1995). This is a disconcerting finding given the marked theoretical and empirical advances that have occurred across all domains of basic psychopathology research and applied clinical assessment over the last 40 years.

Regardless, it remains difficult to select an appropriate tool to assess PDs. At present, there are no best practice guidelines for PD assessment. In 2005, Widiger and Samuel began a related initiative. Based on their review of the extant literature they presented an "evidence-based assessment" protocol for PDs (p. 278). In sum, they recommend use of both a comprehensive broad-band self-report measure that assesses across all PDs, followed by a semi-structured interview to target flagged areas of concern in greater detail. Further, Widiger and Coker (2002) note that given that there are 10 possible *DSM* PD diagnoses, even a two hour, seemingly comprehensive interview based assessment protocol permits "only 90 seconds" to review any given discrete criteria for a single PD (p. 408). As a result, clinicians typically fail to adequately assess the full range of PDs. Thus, a broad-band self-report screening tool can be exceedingly useful for clinicians. With respect to selecting a given tool, however, narrative and quantitative reviews of existing measures report mixed conclusions. Depending on the respective studies'

design properties and conceptualization of various psychometric methods, reliability and validity estimates within and particularly across measures have been interpreted as comparable to chance (e.g., Clark, Livesley, & Morey, 1997) through to excellent (e.g., Widiger & Coker, 2002; Widiger & Samuel, 2005). An overview follows.

Reliability Evidence

As reviewed by Zimmerman (1994), early reliability estimates "for the presence or absence of any PD" were unacceptably low, but they also relied on the use of unstructured interviews (p. 227). For the *DSM-III* field trials (Spitzer, Forman, & John, 1979) and a follow-up independent investigation (Mellsop, Varghese, Joshua, & Hiscks, 1982), inter-rater reliability estimates (kappa coefficients) were .61 and .41 respectively, for joint interviews with the criterion being any PD diagnosis. Only the Mellsop group reported inter-rater reliability estimates for discrete disorders, and the median kappa coefficient was .23. Early test-retest reliability estimates were also low. Test-retest estimates were .54 for the *DSM-III* field trials (1-3 day time interval) and .44 for a six month time interval in a separate study (Pilkonis, Heape, Ruddy, & Serrao, 1991). Since these early studies, however, use of semistructured interviews and objective self-report measures have improved the reliability of PD assessment.

Zimmerman's (1994) review further illustrates that although estimates continue to vary quite markedly across measures, administrators, and disorders, more recent investigations yield average inter-rater reliability estimates (kappas) for semi/structured interviews in the range of .75 for any PD and .70 for a discrete PD (with the lowest estimates of discrete PDs falling in the .50 range on average). Average test-retest reliability estimates over an interval of less than one week for any PD are .58 and .61 for a discrete PD. Estimates for intervals longer than one week (up to

12 months) for any PD are .57 and .44 for a discrete PD. Of interest, Clark and Harrison (2001) note that when dimensional indexes as opposed to a categorical criterion are used, test-retest reliability estimates are substantially higher (.72 vs. .55, respectively). This finding held for dimensional vs. categorical comparisons on self-report measures as well (.69 vs. .40, respectively). Internal consistency reliability estimates for self-report measures are relatively consistent. Average coefficient alpha estimates are .72. Test-retest reliability estimates for self-report measures are moderate to strong. Like interview data, short-term time intervals for self-report measures yield stronger estimates than long-term intervals (.89 vs. .70, respectively).

Validity Evidence

As of 1994, Zimmerman emphasized that there is actually very limited means to adequately assess PD measurement validity because there is no gold standard of comparison. The available interviews and self-report batteries have poor concordance. The measures, however, assess moderate to markedly disparate content and employ different diagnostic decision rules. Thus, strong convergent validity evidence is unlikely. Study methodology is also highly inconsistent which adds to the poor results. Quality of rater training, blind status, and type of population studied vary considerably across studies. Based on their review of published studies, Widiger and Coker (2002) conclude that convergent validity estimates can be improved by using more structured assessment tools. Based on Clark, Livesley, and Morey's (1997) and Widiger and Coker's reviews, convergent validity estimates (kappas) for semi/structured interview formats are approximately .32 to .37 for any PD and .36 for discrete disorders. Estimates for self-report questionnaires (correlations) are approximately .52 to .57. Estimates between questionnaires and interviews were substantially lower at approximately .27. Results also vary by

individual disorder. Convergence estimates are consistently higher for ANT, AVD, and BOR, and extremely inconsistent for COM, HIS, and NAR (e.g., median r = -.33). Consequently, as of 1997 and based on their review of the literature, Clark et al. upheld the same assertion as Zimmerman: It is difficult to generate an informed opinion on the validity of PD assessment measures.

Overall, Clark and colleagues concluded that the current status of PD assessment is essentially equivalent to chance and furthermore, citing Perry (1992), decry this as scientifically unacceptable. Widiger and Coker's more recent review (2002), however, drew more positive conclusions. They emphasize that convergent evidence improves with increased measurement structure and with dimensional vs. categorical diagnostic assessment (rs = . 47 vs. .20). Lastly, Widiger and Coker note that research has amassed to the level where it is now possible to identify consistent patterns within the seemingly discrepant findings that are often obscured through mean comparisons. These can likely be traced to different conceptualizations of PD employed by the various measures. Thus, more recent reviews suggest that the strength of convergent validity evidence varies by measure and disorder and may be stronger than previously assumed.

Lastly, discriminant validity evidence is also weak and may be even more problematic than the convergent evidence (Clark et al., 1997). Nonetheless, two key findings in this domain are important. First, this area of investigation is largely neglected. The extant studies typically do not assess and/or report discriminant validity evidence. Second, the available results focus on the evidence of excessive comorbidity – which has already been reviewed here. Of interest however, Clark et al. note that the comorbidity evidence is "neither universal nor random" (p. 212). This is important because it emphasizes that while some PD constructs are less clearly differentiated,

some semblance of discrete pathology appears to be consistently discerned. More specifically, the authors note that HIS and SZD, as well as COM and ANT rarely co-occur (Clark et al.) whereas DEP and BOR frequently co-occur (Zanarini et al., 2004). Further, as emphasized by Widiger and Coker (2002) and akin to the comorbidity problem: "The absence of much attention to discriminant validity [in scaling] is a recognition that the diagnostic constructs assessed by these measures do not themselves have compelling discriminant validity" (p. 420). Measures should demonstrate less discrimination between highly comorbid PDs and greater between less comorbid disorders. Indeed, this pattern is often identified (Widiger & Coker), but not always (Oldham et al., 1992). Thus, even the harshest critics who conclude that the validity evidence for PD assessment is weak, at minimum, likely need to concede that although less than desirable, the evidence is not nonexistent.

Moreover, convergent and discriminant validity estimates as well as, internal consistency and temporal stability reliability estimates for more trait-based self-report instruments that incorporate both *DSM* and non*DSM* indexes of maladaptive personality concepts are satisfactory to excellent. In particular, the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ; Livesley & Jackson, in press) and the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993a) report the following reliability and validity evidence: coefficient alphas range from .71 to .93, test-retest *r*s range from .68 to .93 (Clark, 1993b; Schroeder, Wormworth, & Livesley, 1992); and subscale correlations between the DAPP-BQ and the SNAP are in the expected direction (e.g., convergent mean *rs* = .5 vs. discriminant mean *r*s = .2; Clark, Livesley, Schroeder, & Irish, 1996), and regression and canonical correlation analyses yield discriminant and convergent evidence in the expected directions across the DAPP-BQ scales with the NEO-PI-R domains and facets (Schroeder,

Wormworth, & Livesley, 2002). Thus, it appears that conscientiously constructed self-report scales of PD related constructs can indeed provide reasonable, psychometrically sound assessments of PD.

Validity Threats

A frequently cited validity threat in normative and disordered personality assessment involves consideration of state versus trait effects. By definition, both personality and PD must demonstrate a degree of stability over time, but the requisite magnitude and duration are debated. In assessment of PD, it seems reasonable that practitioners should have a degree of confidence when administering a self-report measure of PD that the scales are indeed tapping aspects of personality functioning, rather than a purely state artifact (e.g., negative affect). Two lines of evidence address these concerns: (a) stability of PD assessment over time and (b) state effects at time of assessment. Each will be discussed in turn.

Stability of PD assessment over time. A greater wealth of information is available on the stability of normative personality traits over time. Reviews of longitudinal studies suggest that stability coefficients average .75 for a one year follow-up period and .62 in long-term (up to 20 year) follow-up investigations (Clark et al., 1997). As well, normative trait stability estimates increase from childhood (.31) through to early/mid-adulthood (.64) and remain relatively stable by late adulthood (.74) (Roberts & DelVecchio, 2000). By definition, PD is also a relatively stable condition. Indeed, reviews indicate that PDs are relatively chronic, have childhood precursors, and commonly onset during adolescence, but differentially attenuate or remit with age. Variance attributable to biological/genetic factors as well as shared variance has been

demonstrated within and across childhood temperament, adult personality facets, and PD traits (Paris, 2003; Shiner, 2005). As described by Paris and Shiner, course is neither consistent nor linear with monotonic, quantitative increases in symptoms. Rather, course is more typically irregular with periods of both wellness and exacerbation of symptoms. As well, chronicity of any given PD is less stable than chronicity of PD in general. Moreover, even though major symptoms may remit with age (e.g., cessation of criminal behavior among ANTs), personality dysfunction in related domains may not remit (e.g., ongoing marital and employment difficulties; premature death). Chronicity also varies by disorder, with improvement more likely in Cluster B disorders compared to Clusters A or C. Unfortunately, comprehensive, methodologically sound longitudinal data is not available for all of the PDs. Relevant findings of key studies follows.

BOR is one of the more well-studied PDs. Longitudinal investigations have found that approximately 75% of BORs improve after 15 years and 90% by 30 years (Paris, Brown, & Nowlis, 1987; Paris & Zweig-Frank, 2001). Studies of SZD suggest that SZD tends to have a more unremitting course with chronic, pervasive low functioning. Clinically significant recovery has not been demonstrated in 10 year follow-up studies (McGlashan, 1986; Seivewright, Tyrer, & Johnson, 2002), and symptoms can become worse – even with drug and psychotherapy interventions (Seivewright et al.). A review a longitudinal studies by Dolan-Sewell and colleagues (2001) found that premorbid normative and maladaptive personality functioning related to SZD and ANT appears relatively stable over a 10 year follow-up period and uniquely predicts later onset of various Axis I conditions. In a large-scale, multi-site two year follow-up longitudinal study (n = 549), Morey et al. (2004) reported test-retest reliability estimates for the criteria sets of four PDs (BOR, AVD, COM, SZD). Median reliability coefficients based on semi-structured interviews by blind, expert raters ranged from .65 to .84 across the PDs. Also,

through comparing various forms of residualized change scores within and across the disorders, the authors found that change over time was more consistent with disorder specific symptom change rather than an index of generic or global change. Internal consistency reliability estimates also remained relatively stable over time (baseline α = .69 to .83 and .66 to .78 at two year follow-up).

Lastly, within a developmental context, reviews of longitudinal investigations (Johnson et al., 2000) suggest that prevalence of PD traits, on average, moderately declines through adolescence and young adulthood. However, average PD trait stability estimates for a two-year interval during the same age periods are virtually equivalent (.69 and .66, respectively). In general, mean estimates are again somewhat misleading because symptom prevalence and stability varies by disorder (e.g., COM traits do not appear to decline during adolescence). Also as with adults, as the follow-up interval increases, PD trait estimates become more stable than categorical diagnoses. As a cautionary statement, it has been acknowledged by Johnston et al. and others that stability estimates may under-represent true stability over time because many baseline measures in longitudinal studies utilized older assessment tools or unstructured interviews because more reliable tools were not yet available. As well, reliability of measurement including: scale selection, employment of different scales at the various time intervals, disparate rater qualifications, and blind status are all noted potential confounds that can reduce precision and, in turn, stability estimates (O'Boyle & Self, 1990; Zimmerman, 1994). At this time, a tentative conclusion appears to be that the research suggests PDs demonstrate a moderate degree of short-term stability, longer-term stability tends to be weaker and varies by disorder, and estimates are likely somewhat attenuated due to moderately reliable measurement tools and research strategies.

State effects at the time of PD assessment. Assessment validity and particularly test-retest reliability estimates are also believed to be confounded by Axis I pathology. PD measures are often first administered when individuals present for treatment. Depending on the setting, this can be synonymous with some form of acute distress. Kurtz and Morey (2001) conducted a study using the PAI and were the first to assess the accuracy of self-report measures of BOR among individuals currently experiencing a major depressive episode (MDE). They compared a small treatment seeking sample (n = 45, all MDE, 50% comorbid BOR, 50% no BOR) with a matched community control group (n = 20, no history of mental illness). At the time of the assessment the two clinical groups were depressed. Based on SCID, BDI, and PAI-DEP scale scores, there were no mean differences in depression symptom severity across the MDE groups regardless of BOR status. During the MDE, two self-report measures of borderline symptoms, the PDQ and the PAI BOR, reliably distinguished between the three groups. Also, the validity indexes of the PAI indicated that although there was a trend for BOR patients to exaggerate symptoms or complain of extreme distress, there was no indication that the BOR patients were more prone to outright feign illness or respond haphazardly on the self-report measure. Overall, the authors concluded that "the validity of self-report assessments for the diagnosis of BOR is not compromised during episodes of major depression" (p. 298). The authors' results are definitely supportive of this conclusion, but replication with a larger sample and follow-up testing is necessary to place more confidence in their interpretation.

Several researchers have emphasized that it is difficult to assess any PD when individuals are experiencing acute distress related to Axis I psychopathology (e.g., Widiger & Coker, 2002). The available research that has experimentally examined this question appears to have focused

on PD with depression and anxiety. Results indicate that depressed mood can indeed distort patients' responses to PD self-report scales (Hirschfeld et al., 1983, 1989). Three studies by Piersma (1985, 1987, 1989) examined test-retest estimates on the MCMI and MCMI-II in three inpatient samples. This data provides an index of the sensitivity of PD self-report measures to state distress (among other reliability, validity, and measurement error factors). Unfortunately, Piersma did not always clearly describe the nature of participants' presenting concerns. However, all participants' reportedly presented in distress, and the primary concern was most likely an MDE. Results indicated that MCMI derived PD diagnoses were not stable over time (mean hospitalization duration was 20-35 days). Concordance between the MCMI at pre and posttest and with clinician interview was poor. Kappa estimates averaged .21 across disorders for the MCMI pre-post and averaged .11 between the MCMI and interview diagnoses at pretest and .14 at posttest. Additional studies with the MCMI-II report similar findings. Moreover, the pattern of symptom change was not consistent across scales, but scores across all of the MCMI-II PD scales decreased over the course of admission.

Piersma (1985, 1987, 1989) concludes that the series of studies with the MCMI/II indicate that although the PD scale scores change over time, they are more stable than the MCMI/II Axis I related symptom scales and less stable than the MCMI/II normative personality trait scales. This is largely consistent with extant theory of personality and psychopathology. However, although this general trend was indicated, statistical comparisons were not run. This interpretation was based on eyeballing the data. When I computed averages of the stability estimates across the three studies, the results suggest that the basic personality trait domains are demonstrably more stable over time (.65), but the PD and Axis I scales evidenced an identical degree of stability over time (.53 and .53, respectively). Thus, Piersma's results appear to suggest

that MCMI/II based PD scores either lack construct validity or may be susceptible to state effects given that they do not appear to be any more stable than MCMI/II Axis I scores.

McMahon, Flynn, and Davidson (1985) compared a substance abuse sample at intake and again on two follow-up occasions during active treatment (one and three months post-intake). They found that MCMI PD scores dropped from Time1 to Time2, but remained relatively stable during treatment (Time2 to Time3). McMahon et al.'s results suggest that MCMI PD diagnoses become markedly more stable after acute crisis. Thus, it appears that state effects can detrimentally impact self-report PD assessment or, more specifically, the MCMI PD scales may be susceptible to state effects if distress is acute. However, different results are obtained when alternative measures are employed. In a large scale (n = 544), multi-site, two year longitudinal follow-up investigation of the relation between anxiety, depression, and four PDs (with assessments at baseline, 6, 12, and 24 months), Shea and colleagues (2004) found that mood and anxiety disorders were differentially related to BOR, AVD, SZD, and COM. Further, state effects at baseline assessment could not wholly account for the relation between the PDs and Axis I conditions. These authors employed multiple assessment indicators at baseline to determine PD diagnosis and used a semistructured rating scale at each follow-up occasion (DIPD-IV).

Unfortunately, it is difficult to draw firm conclusions across studies because respective authors have made strong statements like, the "state biasing effect…is well known" (p. 2), but then the research cited to support this conclusion is actually mixed (e.g., Ottosson, Grann, & Kullgren, 2000; Reich, Noyes, Coryell, & O'Gorman, 1986). For example, several studies that have assessed PD via self-report and/or semistructured interviews have found either no or only minor influence of mood or anxiety states on PD assessment (Loranger, et al., 1991; Reich et al.,

1986; Trull & Goodwin, 1993). And, as Trull and Goodwin suggest, studies that have attributed changes in PD scores to state effects of mood/anxiety typically have not directly assessed this comparison. Specifically, authors report something akin to a significant main effect for a "time" variable, and then draw the respective inference that the significant finding is attributable to mood state effects. As Trull and Goodwin discuss, identification of the detrimental impact of mood state effects would be more convincing if a clear and consistent association was directly assessed and replicated. As an additional complicating variable, it has also been demonstrated that premorbid personality traits are predictive of later depression, and elevated premorbid PD traits do not wholly remit after recovery from an MDE (e.g., Hirschfeld et al., 1989; Widiger & Anderson, 2003, for a review).

Although state effects likely colour current PD self-reports, it is more conceivable that symptom exaggeration rather than random distortion may influence self-ratings. Indeed, although Trull and Goodwin (1993) documented significant changes in PD scores over time, actual change in number of PD symptoms endorsed or mean scale scores were minimal. Similarly, although Reich et al. (1986) reported a statistically significant change in posttest personality scores when anxiety symptoms had improved, a significant change was found on only 5 of 13 scales. Moreover, the significant changes were only indicated in the participants for whom a six week pharmacotherapy intervention led to symptom improvement. Self-report personality test scores remained stable over the six week period for participants for whom symptoms did not improve (with drug or placebo).

In their review, Clark et al. (1997) also emphasize that emotionality or affect, personality traits, and PD traits are overlapping and interrelated constructs and demonstrate differential stability over time. More affectively laden personality and PD traits are more susceptible to

transient effects or "are more influenced by state affect" (e.g., dependency vs. impulsivity) (p. 222). Overall, it remains difficult to tease apart state vs. trait effects particularly when the debate is couched within the larger context of fuzzy boundary conditions between mood, anxiety, personality, and PD constructs. As highlighted by others, it is a practical reality that assessment of Axis II routinely occurs when individuals present in some form of distress. In turn, state effects will likely pose some degree of challenge to accurate self-report PD assessment. It does appear, however, that the degree of influence is not necessarily invalidating. Notwithstanding, empirical study is not needed to recognize that self-report PD assessment is not reliable or valid if conducted when an individual is acutely ill (e.g., manic, psychotic). It is assumed that an ethical practitioner would not attempt to administer a self-report scale of PD under these circumstances.

Personality Assessment Inventory

Overview

Although the survey data previously reviewed indicates that the MMPI/2 remains the most widely endorsed broad-band measure of psychopathology by practitioners, as reviewed by Piotrowski (2000), the PAI has gained acceptance among forensic practitioners (Boccaccini & Brodsky, 1999; White, 1996) and is endorsed by both academic and internship training programs (Belter & Piotrowski, 2001; Piotrowski & Belter, 1999). Thus, interest in and endorsement of the PAI appears to be growing. Because all PAI scales were derived through a rational-empirical strategy, it may prove a viable alternative to the MMPI-2 which has weaker psychometric properties given that the MMPI-2 clinical scales remain empirically keyed. The psychometric limitations of purely empirical approaches to scale construction have been acknowledged since at

least 1957 (Loevinger). As a brief overview, the PAI was published in 1991 and is described as an "objective inventory of adult personality" (Morey, 1991, p. 1). It is a broad-band self-report battery with 22 full scales that subsume 43 subtests[6]. It is intended to "provide information relevant to clinical diagnosis, treatment planning, and screening for psychopathology" (Morey, p. 5). Of interest however, the PAI only specifically assesses two discrete personality disorders (Borderline and Antisocial) and two bipolar personality trait dimensions derived from interpersonal theory (Dominance and Warmth). Thus, only four of the 22 full scales are directly related to personality function. Hence, despite being titled a personality assessment measure, given the content areas encompassed by the tool, a more apt characterization is a broad-band measure of adult psychopathology as opposed to personality.

Psychometric Properties of the PAI

Morey (1991) never explicitly states the theoretical foundation of the PAI. A review of the manual indicates that a variety of domains were consulted, but the *DSM* model appears the most strongly represented. All scales were created through a rational-empirical approach, and initial item inclusion was reportedly determined through consideration of scholarly literature, *DSM* constructs, alternative diagnostic manuals, and clinical experience. Morey's research team created over 2200 original items which were subsequently reduced to 1086 through a process of expert rating on quality and appropriateness, a bias review, and an expert sort. Items were ultimately retained based on an assessment of their empirical properties following pilot testing among university student, community, and clinical populations. Both classical and item response theory approaches were employed to assess item and scale psychometric properties. No single criterion or formula determined final inclusion. Reportedly, items that performed the best across

[6] A listing of the PAI full and subscale names is included in Appendix A. Item content is listed in Appendix B.

all of the psychometric indexes assessed were retained in the final version (final item $n = 344$). From a scale construction standpoint, reported strengths of the PAI identified by researchers and practitioners (Rogers et al., 1998; Trull, 1995; White, 1996; Widiger & Coker, 2002) have included the following: non-overlapping scales, low literacy demand, Likert response format, facet level assessment of PDs, provision of norms for extreme clinical elevations, comprehensive assessment of relevant domains of psychopathology, inclusion of validity/response bias scales, a rational-empirical construction approach with due attention to content validity (e.g., comprehensive coverage of all facets of the respective disorders), ease of computer scoring and profile interpretation, and, overall, the PAI has been described "efficient, inexpensive, and accurate" (White, p. 38).

Reliability of the PAI. Reliability estimates derived in the norming process and reported in the PAI manual follow. Internal consistency reliability estimates for the clinical scales were moderate to strong across both full and subscales. Coefficient alphas for the full scales across the census, college, and clinical norm groups ranged from .66 to .94 with a mean of .86. Coefficient alphas for subscales were slightly lower and ranged from .51 to .89 with a mean of .78. Test-retest reliability estimates were provided only for the nonclinical groups. Stability estimates were strong for both sub- and total-scale scores over an approximate 25 day interval, with a range from .68 to .85. In addition, absolute change in T scores were reported as an alternative index of stability. Mean absolute differences in T scores were minimal and ranged from 2.8 to 4.9 over the approximate 25 day time interval. Boyle and Lennon (1994) also ran a similar test-retest trial over a 28 day interval with a nonclinical sample and yielded similar findings (median $r = .73$,

range .62 to .86). Note that although IRT based item analyses were also run, to my knowledge, the results have never been published.

Results of the initial structural reliability estimates run by Morey (1991) were mixed. Several of the PAI full and subscales were moderate to strongly correlated (e.g., on average, the DEP, ANX, ARD, and BOR[7] scales were correlated .69 across the clinical and community normative samples). Morey also ran a principle components analysis (PCA) and attempted to interpret a four-factor solution. However, inspection of the reported data does not appear to support his initial conclusion (e.g., too many dual loading items), and it is questionable whether a PCA with orthogonal rotation is an appropriate method for this data set (e.g., scales are strongly correlated). Because the PAI scales do not contain overlapping items, Morey interprets the high inter-scale correlations as a reflection of true symptom overlap within and across the domains of Axis I and II psychopathology. Morey also ran confirmatory factor analyses. Reported fit indexes were strong, suggesting that the intended full and subscale structure of the PAI is supported in a clinical population (n = 1246). However, Morey did not provide sufficient information regarding the respective fit indexes to independently evaluate this conclusion. Based on the data provided in the manual, the preliminary findings suggests that the scale structure of the PAI is moderate to strongly supported.

Independent reliability investigations have similarly yielded positive findings. In an inpatient sample (n = 111) with a primary diagnosis of either mood or psychotic spectrum illness, Boone (1998) found excellent internal consistency reliability estimates for the clinical and interpersonal full scale scores. All coefficient alphas were over .77. Reliability estimates were less strong and more variable on the subscale indexes. Alphas ranged from .42 to .88 with a

[7] DEP = Depression scale, ANX = Anxiety scale, ARD = Anxiety-Related Disorders scale, and BOR = Borderline Features scale

mean of .66. In a treatment seeking eating disorder population (n = 238) with a subsequently confirmed diagnosis of Anorexia, Bulimia, or Binge Eating Disorder, internal consistency reliability estimates were again strong for the clinical and interpersonal full scales (M = .82, range = .75 to .93) and moderate for the respective subscales (M = .74, range = .53 to .91) (Tasca, Wood, Demindenko, & Bissada, 2002). In a large university student sample (n = 1697), Trull (1995) also reported strong evidence of internal consistency (α = .84) for the PAI BOR scale, which was demonstrably superior to two concurrently administered measures: The borderline scales of the Personality Diagnostic Questionnaire (PDQ) (α = .54) and the MMPI (α = .67). In a sample of voluntary, inpatient substance abuse treatment program participants (n = 185), internal consistency reliability estimates for the full clinical scales were again strong (M = .87, range = .75 to .92). Also similar to the previous studies, estimates for the subscales were moderate (M = .76, range = .59 to .89) (Schinka, 1995). Similarly, in a mixed sample of clinical (schizophrenia or alcoholism, n = 60) and nonclinical adults (college and community sample, n = 151), Boyle and Lennon (1994) also found moderate to strong estimates of internal consistency (median α = .83). However, rather than a positive finding, Boyle and Lennon raised the issue that such high estimates may conceivably indicate "narrow scales with excessive item redundancy" (p. 182). This has been referred to as the attenuation paradox (Loevinger, 1954). However, given the range of moderate to strong internal consistency estimates obtained across the various studies and samples, it appears more reasonable to conclude that the estimates obtained are more consistent with a desirable degree of measurement precision that is seemingly congruent with the intended unidimensionality of the scales and the respective latent constructs.

Independent investigations of test-retest reliability of the PAI scales is scant. Two studies by Trull and colleagues assessed (among other variables) test-retest estimates for the BOR scale

in an undergraduate sample over a 2 to 12 week time interval. Results of simple test-retest correlations were encouraging and demonstrated moderate stability estimates (*r*s = .73 and .77) (Trull, 1995; Trull, Useda, Conforti, & Doan, 1997). However, BOR score stability estimates appear markedly weaker when examined from the perspective of criterion group comparisons. Specifically, using T scores of 70 as the criterion in both studies, of the groups of 103 and 119 students who completed pre- and post-test BOR scales, 72% and 74% respectively, retained the same classification. Consistent with regression to the mean effects, the more extreme groups experienced the largest categorical change: Of the participants in the two symptomatic groups, 40% and 47% dropped below a T score of 70 at retest, whereas only 9% and 2% of the participants in the two normative groups subsequently exceeded a T score of 70 at retest. Unfortunately, the authors did not provide detailed scale score information, so it is not possible to identify more dimensional information related to the degree of change. For example, a change in scale score from 70 to 69 would result in a significant change in the criterion group membership, but may not be clinically meaningful. Examination of additional studies is obviously desirable, however, to my knowledge no test-retest data is available for the complete PAI battery within a clinical sample.

Independent investigations of the structural reliability or higher order factor structure of the PAI have yielded weaker support for the scale structure of the PAI than Morey's (1991) original results. Findings are somewhat inconsistent across investigators, populations, and methods of statistical analysis (Boyle & Lennon, 1994; Deisinger, 1995; Schinka, 1995; Tasca et al., 2002). For example, Boyle and Lennon reanalyzed the correlation matrix from Morey's original (1991) nonclinical standardization data set using a different statistical package to rerun Morey's confirmatory factor analyses. The resulting fit indexes were poor and, therefore, did not

support the higher order factor structure suggested by Morey. However, Boyle and Lennon ran their analyses on the factor structure suggested by the results of the exploratory PCA. Although it is somewhat unclear, it appears that Morey ran his analyses on the overt scale structure of the PAI rather than the hypothesized latent, higher-order factor structure. Hence, the results remain inconclusive.

In a large, nonclinical sample (n = 4682), Jackson and Trull (2001) applied both confirmatory factor analyses (CFA) and PCA in an attempt to replicate Morey's (1991) proposed four-factor subscale structure for the Borderline Features scale (BOR). The CFA model fit indexes for Morey's subscale structure were poor (e.g., Comparative Fit Index (CFI) = .74). Changing various parameters of the model (items and factors) only minimally improved the fit (CFI = .86). Results of the exploratory analyses were more consistent with either a two or six factor solution, but the CFA fit indexes were again only moderate. This investigation also identified a few problematic BOR items that could not be easily reconciled (e.g., redundant, double-barreled). Nonetheless, several items loaded on factors consistent with Morey's subscales. Overall, the Jackson and Trull findings suggest that the BOR subscales may assess more than four dimensions which are all theoretically consistent with BOR, but the BOR measure may simply lack sufficient items to adequately scale each additional facet.

A notable exception to the discrepancies found across the factor analytic studies of the PAI, however, is the repeated finding that regardless of the type of factor solution obtained, the first factor identified accounts for the vast majority of variance (>30%, remainder <13% per factor) and appears most aptly characterized as a general index of emotional distress and cry for help. In addition, although the specific order and loading values differ across studies, a second factor commonly elicited appears to characterize individuals with predominant externalizing

symptoms that encompass manic, aggressive/dominant, and antisocial/psychopathic behavior. Overall, internal consistency reliability estimates appear moderate to strong and hold across populations. Preliminary test-retest evidence is moderate, but there is an insufficient number of investigations to draw well-informed conclusions. Structural reliability estimates are weaker and vary across populations. Lack of consensus on the exact number and the composition the respective factors for the PAI in the independent structural investigations is likely attributable to problematic statistical applications in the factor and PCA methods applied. In particular, sole reliance on scree plot results or the "eigenvalue greater than one" rule to determine number of factors and the use of factor based techniques with samples of substantially less than 500 were commonly employed despite measurement limitations in these approaches (Tabachnick & Fidell, 1996; Zwick & Velicer, 1986).

Validity of the PAI. Validity estimates derived in the norming process and reported in the PAI manual were extensive in terms of sheer number of correlations derived across a host of measures. However, the samples were relatively small, hence, the data are acknowledged to be preliminary. The comparison samples' composition and size varied across the type of measures administered. The largest clinical comparison group (n = 235) predominantly consisted of voluntary inpatients (61% male) with a drug or alcohol disorder (42%). Of those remaining, 25% had a mood disorder, 3% a PD, 3% an anxiety disorder, 9% an adjustment disorder, and 6% schizophrenia. Convergent and discriminant validity evidence was strong. Correlations with respective subscales of the MMPI, (content, clinical, and PD scales), BDI[8], BAI, BHS, STAI,

[8] BDI = Beck Depression Inventory; BAI = Beck Anxiety Inventory; BHS = Beck Hopelessness Scale; STAI = State-Trait Anxiety Inventory; IAS-R = Interpersonal Adjectives Scale-Revised; MAST = Michigan Alcoholism Screening Test; DAST = Drug Abuse Screening Test; Bell Inventory = Bell Object Relations Inventory; Hare Scale = Self-Report Psychopathy test.

IAS-R, NEO-PI, MAST, DAST, Bell Inventory, and the Hare Scale were in the expected directions. Of particular interest here, the Borderline scale of the PAI (BOR) correlated .7 with the MMPI-PD Borderline scale, .67 with the NEO-PI Neuroticism scale, and .5 with the MMPI-PD Antisocial scale, whereas the PAI Antisocial scale (ANT) correlated .7 with the MMPI-PD Antisocial scale, .82 with the Hare Scale, .4 with the MMPI-PD Borderline scale, and .15 with the NEO-PI Neuroticism scale. The PAI interpersonal scales (IP) correlated in the expected directions with the IAS-R Dominance and Love vector scales, but did not discriminate well across any of the MMPI-PD scales. The PAI Dominance scale correlated .55 with the NEO-PI Extraversion and .36 with the Conscientiousness domain scores. The PAI Warmth scale was moderately, positively correlated with the NEO-PI Extraversion, Openness, and Agreeableness scales at .45, .30, .45, respectively. Overall, preliminary convergent and validity evidence for the PAI is highly encouraging and appears moderate to strong across the various comparisons.

In the 15 years since the publication of the PAI, independent validity investigations have similarly yielded positive findings. For example, in a sample of students previously exposed to a traumatic event (n = 140), the PAI outperformed the MMPI-2 in correctly differentiating groups by presenting problem: depression, social phobia, or posttraumatic stress disorder (Mcdevitt-Murphy, 2004). In an inpatient sample (n = 24) with predominantly psychotic spectrum illness, Klonsky (2004) found that the PAI SCZ scale outperformed the Rorschach's Schizophrenia Index across all classification indexes assessed, including sensitivity (78% vs. 44%), specificity (75% vs. 60%), positive (70% vs. 40%) and negative predictive power (86% vs. 64%), and overall diagnostic accuracy (79% vs. 54%), respectively. In a forensic sample, Rogers, Ustad, and Salekin (1998) found moderate to strong convergent and discriminant validity evidence for the PAI against the Schedule of Affective Disorders and Schizophrenia (SADS) and the Suicide

Probability Scale (SPS). Rogers et al. also demonstrated that the PAI SUI scale provides incremental validity over the PAI DEP scale (additional 24% of the variance) in predicting SPS suicide scale scores. In a college sample (n = 200), Ruiz, Dickinson, and Pincus (2002) found that the PAI ALC scale demonstrated evidence of criterion validity with the SCID-I (M sensitivity = 93%, M specificity = 36% for T = 70), which supports its utility as a screen for alcohol abuse and/or dependency concerns in a nonclinical population. Other research teams have also reported supportive findings for discriminant and convergent validity for both the ALC and the DRG scales of the PAI in forensic populations (Parker, Daleiden, & Simpson, 1999), community and active using populations (Kellogg, et al., 2002), and inpatient treatment seeking populations (Alterman, et al. 1995).

Few studies appear to have examined the PAI specifically in relation to personality functioning - whether normal or PD, and the majority that have are forensic investigations. For example, in a combined outpatient and secure custody forensic sample (n = 127) Douglas, Hart, and Kropp (2001) found that criminals with a history of violent offending, Axis I psychiatric disorder, or PD scored significantly higher (large effect sizes) on the respective PAI scales than criminals without such history. Moreover, in logistic regression analyses, discrete subscales were often the strongest indicators. For example, the SCZ, PAR, and MAN total scales did not significantly predict psychotic spectrum illness among inmates, but the SCZ-S (social detachment) and MAN-G (grandiosity) were significant predictors. Also, BOR and ANT scale scores were significantly higher in the forensic participants with PD compared to those without PD, but only the BOR-A (affective instability) uniquely predicted PD.

In a combined forensic psychiatric and sex offender sample, Edens, Hart, Johnson, Johnson, and Olver (2000) demonstrated moderate convergent validity evidence for the PAI

ANT scales with the Hare Psychopathy Checklist (Revised, $r = .40$ and Screening Version, $r = .54$). The authors note, however, that the ANT scale/s performed well as dimensional indicators of psychopathy, but were less effective for categorical diagnoses. More specifically, the ANT scales correlated most strongly with the behavioral and antisocial lifestyle components of psychopathy as opposed to the interpersonal and affective components (e.g., callousness, lack of empathy). The authors describe this finding as a common limitation of self-report assessment of psychopathy. In a female inmate sample ($n = 78$) Salekin, Roger, Ustad, and Sewell (1998) similarly demonstrated convergent evidence for the ANT scales with the PCL-R, however, they found that the ANT-E (egocentricity) and AGG-V (verbal aggression) subscales uniquely contributed to the prediction of recidivism.

In nonforensic investigations, satisfactory validity evidence for the interpersonal and PD based PAI scales has been demonstrated. Costa and McCrae (1992b) reported moderate to strong evidence of convergent and discriminant validity for the PAI with the NEO-PI personality domains and the psychopathology scales of the Basic Personality Inventory (BPI) in an adult community sample ($n = 117$). In a small treatment seeking sample ($n = 65$), Kurtz and Morey (1998, 2001) demonstrated convergent validity evidence for the PAI BOR scales with the borderline scales of the Diagnostic Interview for Personality Disorders (.78) and the PDQ (.90). Among individuals diagnosed with BOR based on structured interview (DIPD) results, BOR scores were significantly higher among the borderline group compared to both community and nonBOR psychiatric controls. In comparing an inpatient vs. undergraduate sample, Bell-Pringle, Pate, and Brown (1997) found that the PAI BOR scale compared to an MMPI-2 clinical scale 3-point code was better able to correctly classify individuals with BOR diagnosis (82% vs. 9%, respectively). In a large university student sample ($n = 1697$), Trull (1995) also demonstrated the

utility of the PAI BOR scales in diagnosing BOR. The BOR demonstrated strong evidence of convergent validity with the borderline scales of the PDQ (.68) and MMPI (.62). As well, elevated BOR scores were significantly correlated in the expected directions with respective scales of the BDI, IDD, PANAS-X, NEO-PI, PCS, BSI[9], and the borderline criteria from the Structured Interview for DSM-III-R Personality (SIDP-R). In the same student population, elevated BOR scores also uniquely predicted negative outcomes across several domains including: interpersonal functioning, academic success, and Axis I comorbidity over a two year follow-up period (Trull et al., 1997). Also, 13% of the students with elevated BOR scales satisfied diagnostic criteria for BOR as determined by the SIDP-R, whereas no students with normative BOR scores satisfied BOR criteria. Thus, as emphasized by Trull (1995) and (Morey, 1991), elevated BOR scores appear indicative of hallmark BOR features and impaired functioning, but an elevation on this scale in and of itself is not necessarily diagnostic of BOR.

STRUCTURAL

Item Response Theory

Overview of Scale Development

In the field of psychological measurement, given the extensive emphasis placed on discussing issues related to item responses on measurement scales, total scale scores, and other psychometric properties of tests, as pointed out by DeVellis (1991), it is easy to forget that the primary goal of researchers and practitioners is to understand the underlying construct that scales are specifically created to measure. Hence, a fundamental task in test development is identifying

[9] BDI = Beck Depression Inventory, IDD = Inventory to Diagnose Depression, PANAS-X = Positive and Negative Affect Scale – Expanded form trait version, NEO-PI = NEO Personality Inventory, PCS = Personal Coping Styles, and the BSI = Brief Symptom Inventory

which of the available means of test construction provides the best estimate of the construct of interest – in this case, personality disorders. Traditionally, in the realm of personality assessment the most popular measurement model for scale construction has been classical test theory (CTT) (Embretson & Reise, 2000). As will be outlined below, however, CTT is not necessarily the most accurate or the most useful model for all measurement applications. IRT is one potentially viable alternative. To facilitate an understanding of the advantages of IRT, a comparison between CTT and IRT from both logical and empirical perspectives is detailed below.

Comparison of IRT versus CTT

Conceptualization of the construct of interest (or latent variable) is obviously an important consideration across all measurement models. In both IRT and CTT qualities of the latent variable are assumed to drive responding to test items. Hence, latent variables are said to "cause the item score" (DeVellis, 1991, p. 13). As further described by DeVellis, given or upholding this causal assumption various theorems or "empirical relationships" can then be inferred (p. 13). One of the key inferences of both measurement models derived from this causal assumption is the proposition that if a latent variable is driving item responding, then items that assess the respective latent variable should have some type of measurable interrelationship. It is at the point, however, at the stage of defining the nature of (the assumed) item interrelationships, that CTT and IRT diverge. Each measurement theory has its own system of assumptions from which various propositions or theorems have been derived that, in turn, inform the various statistical analyses used to answer practical measurement questions. As emphasized by Hambleton and Jones (1993), at the core of this divergence is an operational understanding of two concepts, true test score and latent trait. A comparison of the fundamental assumptions of

each model is outlined in Table 4.

Caveat. IRT is not a unitary theory. Different theorists have devised several models that have been categorized under the rubric of IRT. The most significant division lies between Rasch based modeling (few item analysis parameters) and more complex IRT models (several item analysis parameters). Divisions across the models also reflect analyses devised for dichotomous versus polytomous item responding and for unidimensional versus multidimensional scaling. Several models hold different assumptions under IRT. As a result, the following outline of IRT assumptions is primarily based on the original, Rasch-based conceptualizations. Divergences from this model will be discussed later, with emphasis being placed on modeling techniques that are most appropriate for PD assessment. Note as well that the divisions in the field of IRT are somewhat contemptuous and ongoing. For example, authors submitting studies to the *Journal of Applied Measurement* are requested not to use the term IRT in reference to Rasch modeling.

Table 4

Core Assumptions of CTT and IRT

Core Assumptions	
CTT	**IRT**
•single or multiple traits influence item responding •"observed score is the sum of...true score and error" (Allen & Yen, 1979, p. 56) -relation between true and error scores is additive •the true score is the expected value of the mean of observed scores (infinite repeated testing) •population's average error score is zero •error scores and true scores are uncorrelated •test items fit a linear common factor model or linear conditional probability function -"error of measurement is...unsystematic or random" (Allen & Yen, p. 59) -items are related via latent variable only, not error -"error scores on two different [items] are uncorrelated" (Allen & Yen, p. 58) -"error scores on one [item] are uncorrelated with" another [item]'s true score (Allen & Yen, p. 59) •"individual items are comparable indicators of the underlying construct" (DeVellis, 1991, p. 12) -"the proportion of item variance attributable to the latent variable...is equal for all items" (DeVellis, p. 19) •parallel items have equal true scores and equal error variances -latent variables exert equal influence on all items -extraneous factors exert equal influence on all items	•single or multiple traits influence item responding, and latent trait estimates must then directly correspond as uni- or multi-dimensional •"latent traits can...assume values from $-\infty$ to $+\infty$" (Allen &Yen, p. 240) •"true score has a functional rather than statistical relation to the latent trait..." (Lord, 1968, p. 386) •"if the latent space is one-dimensional, the latent trait...is the same as the true score..., except for the scale of measurement" (Lord, p. 386) •"item characteristic curves have a specified form" (Embretson & Reise, p. 45) •local independence[a] -"for a given...latent trait value, item scores are independent" (Allen & Yen, p. 241) •theta scale (latent trait estimate) does not contain measurement error •"observed score is the true score" (Allen & Yen, p. 240) •"observed score is not equal to the latent-trait value" (Allen & Yen, p. 240) •IRT calibrated items fit a monotonic, increasing function (e.g., normal-ogive function) -probability of a correct item response increases with the strength/severity of the latent construct •measurement error varies with trait level •measurement error varies for each item parameter estimate •specific objectivity: "trait level and item properties ... are estimated separately" (Embretson & Reise, p. 49) •"provides a full model of behavior": person and item parameters are separate, but in same model (Embretson & Reise, p. 49) •"conjoint scaling of trait level and item difficulty": response probabilities change by a constant (Embretson & Reise, p. 49)

Note. (Allen & Yen, 1979; DeVellis, 1991; Embretson & Reise, 2000; Hambleton & Jones, 1993; Lord, 1980, 1986; McDonald, 1999); [a]Alternatively, Lord (1980) posits that local independence is not truly an IRT model assumption because it is a natural consequence of the unidimensionality assumption.

An expanded review of key IRT assumptions from Table 4 follows. If it is not directly obvious from the comparisons outlined above, several fundamental differences exist with respect to the theoretical conceptualization and the practical application of measurement between CTT and IRT. In particular, the assumption of local independence of items and respondents is fundamental to IRT and holds powerful implications for trait estimates: Latent-trait estimates provide essential and sufficient information to predict item or test scores. In theory, for respondents at a circumscribed trait level (i.e., fixed value of the latent trait), knowledge of performance on other test items or norm group membership should not improve the prediction of a given item response or test performance beyond that afforded by the latent trait estimate (Allen & Yen, 1979). Item endorsement and test scores can (and should) vary across the levels of a given trait, but scores should not vary within a specific trait range. The assumption of local independence is, however, further subdivided into weak and strong variants, which highlight more subtle distinctions between CTT and IRT approaches (McDonald, 1999).

As reviewed by McDonald (1999), if the CTT approach of common factor modeling is used in test construction, the respective analyses espouse the assumption of weak local independence. Specifically, the common factor model employs a linear conditional probability function, which yields estimates of "average behavior of [test] items over the range of the test" (p. 249) and assumes pairwise conditional independence. This espouses weak local independence because the assumption explicitly specifies conditional independence at the level of inter-item functioning (e.g., Inter-item "covariances are zero…[at] a fixed value of the factor score(s)", p. 255). Whereas, IRT methods take the assumption one step farther and assume a more rigorous or inclusive stance: The assumption of "strong or full principle" local independence (p. 255) subsumes the principle of inter-item independence (zero covariance) and explicitly specifies the

conditions required to assume a latent trait. Alternatively stated, rather than only referencing zero item covariances at a fixed trait level, the strong approach specifies that "the conditional probability of any pattern of (zero/unit) responses is the product of the conditional probabilities of those responses" (p. 256). As explained by McDonald, in practice, the strong and weak approach translates to the inclusion and exclusion, respectively, of population probability estimates of the item response patterns in the overall statistical model.

More generally, Allen and Yen (1979) explain that, in practice, local independence assumes that "answering one item", "knowledge of one item's answer", and "changing the order of administration of a set of items" should not influence test performance. If this were the case, the scale would not fit a strong IRT model (Allen & Yen). Notwithstanding, to obscure matters somewhat, current IRT modeling has become more complex. As highlighted by Embretson and Reise (2000), IRT models have now been proposed that can accommodate items that are expected to be associated in some way. The identification of the item parameter is simply expanded to include such dependencies.

An additional important comparison detailed by Allen and Yen (1979) is that the definition of true score as conceived within CTT holds no assumptions with respect to construct validity. Under CTT, if measurement error was somehow eliminated, the test score obtained could be accurately described as the respondent's true test score. It is tempting to take this inference one step further and surmise, hence, that the respondent's true test score represents the underlying construct of interest. However, CTT only allows for the inference that a respondent's true score reflects the most accurate test score for a given individual. Determining that the respective true test score (e.g., PAI Depression Scale, $T = 95$) is a valid representation of the

underlying construct of interest (Major Depressive Episode), requires additional validity testing (Allen & Yen, p. 58).

Hambleton and Jones (1993) note additional theoretical differences between the two models. IRT goes beyond the notion of true score and works directly (e.g., theorems) at the level of the latent construct. In CTT, the fundamental component of the measurement theory could be conceived as the true score whereas in IRT, the fundamental component is the latent construct. More specifically, as explained by Hambleton and Jones who credit Lord (1953) for first noting this premise, both "observed and true scores are test dependent" (p. 253): Obtained test scores and estimates of true test scores are a direct function of the type of test administered. Regardless of the stability of the latent construct of interest, at a given point in time, individuals "will have lower true scores on difficult tests and higher true scores on easier tests" (p. 253). Thus, estimates of true scores can vary despite constancy in the underlying latent construct. Therefore, despite the real possibility of obtaining accurate estimates of true scores, tests created within the CTT model can still lead to inaccurate estimates of the latent construct.

As reviewed by Hambleton and Jones (1993), IRT attempts to address this shortcoming by seeking to attain estimates of the latent construct in a manner that is "independent of the particular choice of test items" or in a manner that does not rely on estimating true scores (p. 253). Further, IRT attempts to attain estimates of the latent construct that are sample independent (e.g., invariant person statistics). Nonetheless, although IRT can be conceived as an improvement over CTT in latent trait measurement, IRT methods still require construct validation work. Even though IRT can demonstrate that a tool is accurately assessing a single latent construct, IRT provides no definitive evidence that the latent construct being assessed is the theorized construct (e.g., trait) of interest.

Lord (1980) highlights another fundamental distinction between CTT and IRT. Because IRT is sample independent, scales developed through IRT methodology are stronger predictors of future item responding for any individual. In contrast, the predictive power of CTT based scales is limited to individuals equivalent to those in the normative sample. Moreover, Lord notes that IRT does not assume that each item is an equivalent predictor of the latent construct. IRT proposes that items hold differential predictive power across given levels of a trait. Hence, IRT is capable of providing probability estimates of the effectiveness of a given item across the range of severity of psychopathology.

Limitations of CTT can also be levied from the perspective of generalizability theory (e.g., DeVellis, 1991). As discussed, CTT conceives of item variance as containing true score plus random error variance. Generalizability theory, however, posits that it is more realistic to also consider systematic error variance as a third source of variance in the composition of item scores. Specifically, demographic, individual difference, and method of administration variables, etc. could all conceivably influence item responding and resulting test scores. Generalizability theorists would, thus, suggest that the CTT assumptions are too simplistic because an important, documented source of variance is discounted in applications of CTT theory. Moreover, as an added strength, applications of generalizability theory are capable of identifying sources of problematic error variance. Alternatively, CTT models quantify error, but do not identify the source (Allen & Yen, 1979; DeVellis).

With respect to generalizability theory, as a result of addressing additional sources of error, IRT is an improvement over CTT. As well, if the IRT calibration sample is sufficiently diverse (e.g., adequate representation of individuals at each level of trait across the entire trait dimension), the accuracy or precision of the IRT model and, hence, its generalizability (e.g.,

ICC) is greatly enhanced (Brannic, 2001). IRT does not, however, address all issues relevant to generalizability theory. For example, in IRT, the standard error of measurement (SEM) represents "the informativeness of the items about the person's standing on the latent trait. [In 2PL models, it] depends on both item difficulty and discrimination" (Embretson & Hershberger, 1999, p. 245). Moreover, as discussed by Marcoulides (1999), although IRT has the potential to estimate or explain several additional sources of error through expanding the number of parameters included in IRT models, most researchers fail to do this. However, it is also arguable that "failure to include" is not really a weakness of IRT, but rather a limitation in its application.

Examination of the statistical applications of the abovementioned CTT and IRT model assumptions serves to highlight additional disparities between the two approaches. Yen and Edwardson's (1999) contrasts of common measurement concepts from a CTT and IRT perspective provides a useful means to understand the key differences that result from the application of the divergent assumptions of each model. An expansion on their initiative is outlined in Table 5. Awareness of these distinctions between the models also facilitates understanding how IRT purports to achieve test and sample independence in assessment of a latent construct.

Table 5

CTT versus IRT Measurement Concepts

Measurement Concept	CTT	IRT
Item Difficulty	•proportion of respondents who correctly endorse an item *Common Estimates* •simple fraction or percentage	•probability of a correct item response; varies as a function of level of trait *Common Estimates* *Primary information source:* •location (difficulty) parameter ("b")
Item Discrimination	•inter-relationship between item responses; "difference between the proportion of high...and low-scoring examinees who get the item correct" (Allen & Yen, 1979, p. 122) *Common Estimates* •item-total correlations •point biserial correlations •extreme group comparisons	•"degree...item response varies with ability level" (Lord, 1980, p. 13) *Common Estimates* •"slope of the [ICC] at the inflexion point" (Lord, 1980, p. 13)
Reliability	•"proportion of observed-score variance that is true-score variance" (Allen & Yen, 1979, p. 73) •estimates are positively influenced by test length *Common Estimates* •test-retest correlations (temporal stability) •inter-item correlations (internal consistency) •single estimate of SEM	•precision of measurement •estimates are not dependent on test length *Common Estimates* •item information curves/function •multiple estimates of SEM (separation coefficients; e.g., person and item separation reliability) •differential item functioning
Test Scores/ Interpretation	•total items endorsed •normative comparisons	•optimal model estimate of trait •estimate of relation between test performance and trait •interval scale (easier to achieve vs. CTT) *Common Estimates* •Item response function -trace line, ICC •Test characteristic curve

Note. Allen & Yen, 1979; Bejar, 1983; Hambleton & Jones, 1993; Lord, 1980; Wright & Stone, 1999; Yen & Edwardson, 1999.

Consideration of the measurement concepts in Table 5 reveals the implications of applying the discrepant CTT and IRT measurement models. The fundamental difference between the two models illustrated in the above concepts is the manner through which IRT attempts to achieve sample and item independence. As concluded by Hambleton and Jones (1993), estimates of a latent construct within CTT models are heavily dependent on the respective test given (items) and the testing sample characteristics. For example, a limitation of the CTT conception of item difficulty as the proportion of correct responses is that this statistic is a function of both the items and the sample tested: If the sample is nonclinical, items denoting hallmark PD features are less likely to be endorsed, whereas a clinical PD sample is more likely to endorse such items. Further, if the test has a high or low ceiling, respondent proficiency or degree of pathology may be less than optimally estimated. Because the emphasis of the current project is on IRT, a more detailed discussion of IRT follows and several of the statistical and theoretical differences between the two models will become more apparent.

Theoretical Foundation of IRT

The origins of the IRT perspective reportedly date at least as far back as the early 1900s. Thurstone is one of the earliest measurement experts who is commonly cited for delineating several requisite assumptions necessary for accurate measurement of psychological constructs (e.g., Thurstone, 1926, 1928, 1931; as cited in Wright, 1999). Three of Thurstone's measurement principles of greatest direct relevance for IRT models are the concepts of sample independence, item or test independence, and linearity. Sample independence reflects the principle that in order for a measurement scale to be considered accurate, its calibration cannot be influenced by qualities or properties of the construct to be measured. For example, whether a ruler is being

used to measure the length of a book or the length of a desk, characteristics of neither the book nor the desk affect the calibration of the ruler (e.g., One cm is always calibrated on a linear interval scale at equal intervals of ten mm, regardless of the object being measured). Further, a 25cm book is one-eighth the length of a 200cm table. In a similar way, IRT proposes that an objective scale devised to measure a psychological construct (e.g., PDs) should be calibrated in such a way that properties of the sample do not influence the units of measure. As a crude example, 12 units of borderline features on a PD scale should reflect 12 units of borderline features in any individual that is assessed with the PD scale. And, an individual with 12 units of borderline features should have greater distress or symptomatology than an individual with 6 units of borderline features.

As reviewed by Embretson and Reise (2000), CTT also endorses the value of creating an interval scale, however, CTT based scales achieve interval calibration through seeking a normal distribution of test scores. This process necessitates satisfying two assumptions that are difficult to meet in practice: (a) the true scores must satisfy "interval scale properties", and (b) observed scores must be normally distributed (p. 32). As well, only linear transformations of raw scores preserve the interval calibration (e.g., percentile matching transformations differentially adjust score intervals). Hence, a normal distribution of observed test scores must be obtained for each norm group – a cumbersome process. Therefore, even if the assumptions are satisfied, the CTT interval scale calibration technique still limits the applicability of the scale to the norming group (or equivalent) population. This again highlights the sample dependency concern of CTT based methods. Moreover, as highlighted by Wright (1967), a CTT derived raw scale score does not provide information about which specific items a person endorsed. If the sample is sufficiently large, IRT permits items to be calibrated over a desired range of the construct of interest. As a

consequence, IRT based estimates of the underlying trait are more accurate than raw score values.

As further reviewed by Embretson and Reise (2000), the application of Rasch based IRT models facilitates the creation of interval calibrated scales through means that eliminate the dependency on sample characteristics and normally distributed traits. In essence, the Rasch based models work backward compared to CTT models. As opposed to deriving a model from the obtained data, similar to structural equation modeling, in IRT applications a theoretically derived statistical model specifying characteristics of an interval scale for the target trait is created (e.g., equal intervals across level of trait for endorsement of both easy and difficult items). The obtained data are then compared for fit against the statistical model. If the data fit the model, sufficient evidence for an interval scale is achieved. In theory, the scale will be applicable to any population because the calibration of the scale (e.g., intervals) was determined by a mathematical proof rather than participant response.

In more specific terms, Embretson and Reise (2000) explain that the person variables (latent trait) and item variables (difficulty) "are placed on a *common scale* in IRT models" (p. 128). The probability of correct item endorsement is set to match the respondents' trait level. For example, a respondent with high trait characteristics will have a high probability of success/endorsement (e.g., 98%) with items that tap low levels of a trait (easy items). The probability that the same individual will successfully answer or endorse high trait items (difficult items) will be closer to 50% because the items are testing at threshold for a high trait individual. Hence, item independence reflects the principle that if "item responses are optimally weighted, *the contribution of the item to the measurement effectiveness of the total test does not depend on what other items are included in the test*" (Lord, 1980, p. 22). Moreover, IRT's computation of

item difficulty and latent trait on the same metric has the distinct advantage of enabling test developers to create scales that more accurately discriminate across individuals at any desired trait level (Embreston and Reise).

The application of Rasch based IRT models also uphold the principle of conjoint additivity. Wright (1999) maintains that the principle of conjoint additivity is another "decisive theoretical requirement for measurement" (p. 80). Conjoint additivity reflects the premise that test items should be calibrated in such a way that a correct response to any difficult item necessarily presumes a correct response to all items that are deemed easier. According to Wright, only Rasch models uphold this principle. As others have argued (e.g., Reise, 1999), however, it is inherently more plausible to uphold this principle when applying IRT to the measurement of constructs like math proficiency or vocabulary. In certain domains, innate and/or acquired proficiency or pathology may reflect a more linear, quantitatively increasing construct (e.g., computing algebra necessitates mastery of simple multiplication). In comparison, when diagnosing a given personality disorder, degree or prototypicality of psychopathology may prove to be a less linearly definable construct. Nonetheless, from a logical/theoretical perspective, upholding the principle of conjoint additivity in personality disorder assessment may not be unattainable, simply more challenging in terms of item construction.

IRT Computation. Reduced to its simplest form, the basic statistical premise of IRT is relatively easy to comprehend. The basic formula for IRT modeling is illustrated in Figure 1.

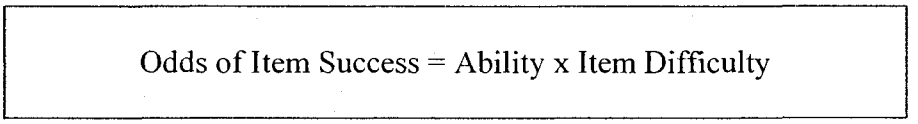

Figure 1. Essential IRT Model (Wright, 1967).

As reviewed by Lord and Novick (1968), all IRT applications are based on additions and augmentations to the essential IRT function: the item characteristic curve (ICC). An example of the simplest IRT function, the one-parameter logistic function for binary items (1PL) is displayed in Figure 2 (cf., two-parameter and three-parameter models described below). The first premise or task in applying IRT is to estimate the latent trait. Applications of IRT achieve this through analyzing response patterns to scale items. The simplest form of ICC represents the regression of "the probability of item success on trait level", and the probability is plotted as "a monotonic and increasing function of trait level" (Embretson & Reise, 2000, p. 46). Meaning, as trait levels increase or decrease, item endorsement must similarly increase or decrease; it cannot oscillate. Although the general shape of the curve resembles an "S" (ogive), the exact shape depends on the type of model specified. Again depending on model specifications, ICCs can also differ in location, slope, and asymptote, reflecting differences in item difficulty, discrimination, and response range restrictions, respectively. The dependent variable is often predicted as a probability or log odds. In the simplest Rasch model, using a log odds approach, the ICCs are linear. This is the most desirable model because it upholds all of the rigorous IRT assumptions which, in turn, yield the strongest inferential power for application of the results (e.g., true interval measurement). A conceptual and statistical representation of the formula for the three-parameter IRT model (3PL) is displayed in Figure 3, and an illustration of the respective item response function/s are outlined in Figure 4. Note that the 3PL model is illustrated here because, as the equation in Figure 3 demonstrates, the 3PL formula easily accommodates an explanation of the 1PL and 2PL models.

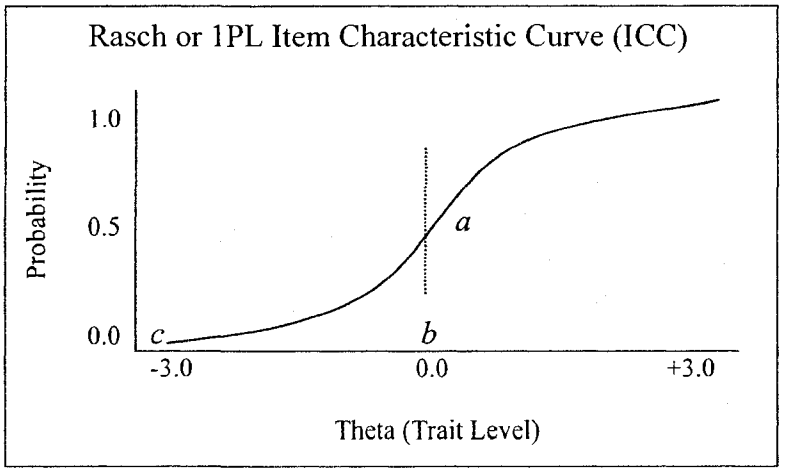

Figure 2. 1PL item characteristic curve: a = slope or discrimination parameter (fixed to a constant in 1PL); b = intercept or difficulty parameter (approximately zero in this example); c = lower asymptote or guessing/pseudo-chance parameter (fixed to zero in 1PL).

3PL Model

$$\text{Probability of item endorsement} = \frac{1}{\text{trait level} \bullet \text{item difficulty} \bullet \text{item discrimination} \bullet \text{guessing}}$$

or

$$P(\theta) = c + (1 - c)\frac{1}{1 + \exp\,(a\,(\theta - b))}$$

Figure 3. Conceptual representation and statistical equation for the 3PL IRT model: θ = latent trait estimate; a = slope = discrimination parameter; b = intercept = difficulty parameter; c = guessing/pseudo-chance parameter. For the 2PL model, c is fixed at zero; and for the 1PL model, both a and c are fixed at zero.

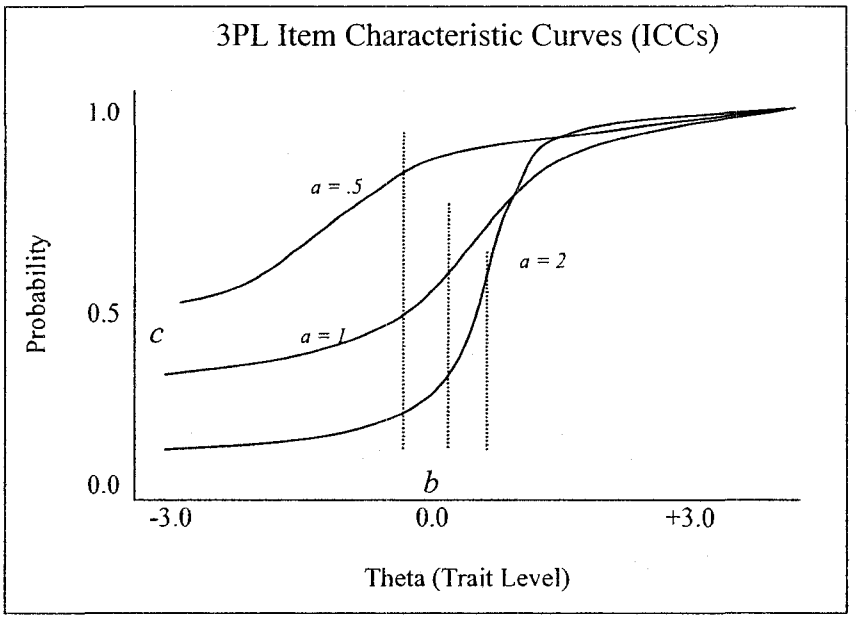

Figure 4. Examples of 3PL item response functions (ICCs) with varying levels of *a*, *b*, and *c* parameters.

It is noteworthy to mention here that Lord, Birnbaum, Embretson, and like-minded IRT theorists subscribe to the above conceptualization of IRT – typically conceived as two-parameter (2PL) or three-parameter (3PL) modeling. However, Rasch, Wright, Fischer and their followers do not support the above assertion and endorse only Rasch based or one-parameter (1PL) IRT models. For example, Wright (1999) outlines several harsh criticisms of 2PL and 3PL IRT models including, failure to demonstrate conjoint additivity, poorer minimization of mean square residuals, lack of additive parameters and sufficient statistics for each parameter, and lack of construct stability owing to permission of crossing ICCs (e.g., differential item functioning). Wright identifies the consequence of his perceived limitations of the 2PL and 3PL IRT models as prohibiting the development of a truly objective, sample and item independent measurement scale. More specifically, the above limitations are conceived as evidence of item bias, construct

multidimensionality, respondent error (e.g., guessing is characterized as "unreliable person lability", p. 98), and overall lack of scale reliability. Ultimately, Wright argues that theorists/researchers who use non-Rash based models have failed to adequately apply the logical and mathematical principles necessary for objective measurement.

In contrast, Lord (1980) argues that Rasch based models are simply a "special case of the three-parameter logistic model" (p. 189). Lord argues that the Rasch models can be limited because they assume that "all items are equally discriminating" and contain no allowances or caveats for potential guessing behavior by respondents (p. 189). Thus, in 3PL model terms, Rasch based models hold the discriminating parameter constant across all items and set the guessing parameter to zero (Lord). This also seems equivalent to the 2PL model, which simply fixes the guessing parameter. Lord further states that the assumptions of the Rasch based models are rarely upheld and argues, therefore, that the model is typically a poor fit for real-world data. Hence, Lord seems to assert that differential item discrimination and respondent guessing are essential measurement considerations that need to be incorporated into accurate, effective estimation procedures. Wright (1999), in defense of Rasch modeling, retorts that incorporation of such parameters yields scales that at best provide "local descriptions of transient data" that have no "inferentially stable meaning" (p. 99). Wright acknowledges that differential item discrimination and respondent guessing are important sources of error to document and find means to overcome. He contends, however, that they cannot be incorporated in a scale development exercise because they preclude construct stability, construct validity, and any sense of linear, interval, and item- or sample-independent scale calibration.

The logical and mathematical underpinning of the Rasch models is more convincing with respect to the capability of developing a truly objective, sample and item independent

measurement scale. The science behind PDs, however, is not yet sufficient to permit development of this type of measurement tool: A ranked (interval or ratio scaled) list of signs or symptoms that uniquely characterize discrete PDs that can be transformed to questionnaire format remains partially unknown. Further, the rigorous assumptions that need to be upheld in Rasch based modeling to ensure objective measurement may be unrealistic for measurement across virtually all domains of psychopathology. For example, in polytomous Rasch based IRT modeling, the probability of endorsing a given ranked response on a Likert-style scale must be assumed independent of the probability of endorsing any other response option (Masters, 2001). The reality of this assumption holding when applied to Likert-style measures of psychopathology (e.g., PAI scale items) seems unlikely.

Selection of IRT Model

As reviewed by several authors (e.g., Embretson & Hershberger, 1999; Hambleton, 1989; Hambleton & Jones, 1993), since its inception, models of IRT have substantially evolved. For example, original conceptualizations only considered models for scales with a dichotomous response format that assessed a single, unidimensional latent trait. Alternatively, several strategies are currently available to conduct IRT modeling. In a brief review, Embretson and Hershberger (1999) note that many psychometrists strongly endorse certain techniques over others, with no consensus across the field. Less than an issue of which modeling technique is psychometrically superior, it might be more relevant, however, to consider which modeling technique is most appropriate for the latent construct and research question of interest. Embretson and Hershberger note three issues to consider in model selection: parsimony, appropriateness/empirical fit for the data, and dimensionality of the construct (single or

multidimensional). Sensitive to these criteria, a reasonable option for the current study is Samejima's Graded Response Model (GRM; 1997), which will later be discussed in greater detail.

Dimensionality. Most IRT models are grounded in the assumption of unidimensionality of the latent construct. Of concern, personality disorder is a multidimensional construct that is not necessarily most aptly captured by a series of unidimensional scales. As previously reviewed, PD researchers have cogently argued both logically and statistically/empirically in support of dimensional models of PD (e.g., O'Connor & Dyce, 1998; Widiger & Frances, 1994). In particular, the five-factor model of normal personality was noted as the dimensional model that held the most promising fit for abnormal personality (O'Connor, 2005a). Nonetheless, because the primary goal of the current study is to generate practically useful PD scales for applied settings and given that the categorical model (*DSM-IV-TR*) is the current gold-standard for PD diagnosis in clinical settings, the categorical model of PD will be applied in the present study. In turn, the psychometric properties of a series of scales intended to assess unidimensional PDs will be examined in the current study. It is expected that the unidimensional assumption will hold because the underlying personality pathology (trait/PD) assumed to drive item responding on a given scale is expected to be a single disorder (e.g., antisocial PD). Notwithstanding, because unidimensionality of the underlying pathology (PD) and adequacy of the *DSM* diagnostic model for PD symptom presentation are simply assumptions, the appropriateness (or fit) of a unidimensional IRT model for each PD scale will need to be explicitly tested in the current study.

Should the unidimensional model assumption prove unsatisfactory, a seemingly obvious solution for assessing a multidimensional construct is to use multidimensional IRT (MIRT) modeling. McDonald (1999) advocates an MIRT model for binary data that blends unidimensional IRT with CTT based principles of exploratory and confirmatory factor analysis. As explained by McDonald, the proposed method is essentially equivalent to running a concurrent series of interconnected, unidimensional IRT models. McDonald's MIRT model, however, is problematic for the purposes of the current study because it is only applicable to dichotomous data; the measurement errors of the interconnected IRT functions include reliable variance; and the model has difficulty accommodating complex items. In the writer's estimate, applying McDonald's MIRT principles to a polytomous data set would be exceedingly complex.

Several researchers have documented that MIRT remains a developing field, and a theoretically sound and practically useful statistical MIRT model for polytomous data does not yet exist (e.g., Kirisci, Hsu, & Yu, 2001; Reckase, 1997). For example, Embretson and Reise (2000) note that adaptations of nonlinear factor analyses to MIRT (e.g., full-information item factor analysis, TESTFACT 2 program; Wilson, Wood, & Gibbons, 1991) indeed overcome many of the short-comings of traditional factor analysis, the method remains only applicable to dichotomous data. As well, more recent Monte Carlo simulation studies have yielded some support for two-parameter MIRT models (e.g., Bolt & Lall, 2003), however, the estimation procedures only assess two-dimensional constructs and personality dysfunction is known to vary across more than two dimensions. Other MIRT models exist that can accommodate more than two dimensions, however, researchers have demonstrated that as the number of dimensions increase, the soundness of the statistical application of the MIRT theorems is increasingly compromised. As reviewed by Reckase (1997), although much progress has been made in the

theoretical and mathematical domain of MIRT model development, the creation of statistical estimation procedures to practically apply the propositions of the respective models to real-world data has substantially lagged behind. As a result, MIRT methods are not considered a viable option for the current project and, hence, are not reviewed here.

Model Fit: Polytomous Model. Because the PAI has a Likert-style response format, the selected IRT model must be able to analyze ordered, polytomous response formats. The Likert-style PAI response format can be considered as either a nominal or ordinal scale (e.g., PAI response options are: false, slightly true, mainly true, very true). As cautioned by Wright (1999), although it is assumed that respondents will interpret and respond to Likert-style scales as if they are a true ordinal scale, this premise needs to be tested. A dichotomous response pattern is equally likely. Thus, although it is anticipated that a polytomous IRT model will be selected, evidence for ordinal properties of the scale will first be examined to ensure that a polytomous model is indeed appropriate.

As reviewed by Embretson and Reise (2000), the fundamental challenge of polytomous models is to account for the range in responding within each item (e.g., PAI has four response categories per item) in addition to the basic IRT parameters. Further, "item discrimination depends on a combination of the slope parameter and the spread of the category thresholds" (p. 112). Each polytomous model incorporates such computations of category thresholds or intersections in different ways. Samejima's GRM is a variant of the traditional 2PL model with an adaptation to accommodate the category thresholds. Specifically, a difference parameter is added, which takes into consideration the cumulative probability of responding to a lower versus higher response option (Baker, Rounds, & Zevon, 2000). (Samejima's GRM formula is detailed

in the results section). The advantages of Samejima's GRM model include: category intersection parameters do not need to be ordered; category intersection parameters can vary across items; and slope parameters can vary across items (Embretson & Reise). Also, Samejima's GRM has demonstrated superior estimates of measurement precision compared to forced dichotomization of a polytomous dataset (Dodd & De Ayala, 1994). Lastly, as a practical advantage, GRM analyses can be run through a freely accessible statistical interface environment ("R" ltm package; Rizopoulos, 2006).

Successful application of Samejima's GRM to the realm of Likert style assessment has been demonstrated across several psychological domains, including assessment of normative personality (Robie, Zickar, & Schmit, 2001), schizophrenia (Long, Harring, Brekke, Test, & Greenberg, 2007), optimism (Rauch, Schweizer, & Moosbrugger, 2008), emotional adjustment (Rubio, Aguado, Hontangas, & Hernandez, 2007), attachment style (Fraley, Waller, & Brennan, 2000), and posttraumatic stress disorder (King, King, Fairbank, Schlenger, & Surface, 1993). Moreover, Samejima's model may be superior to alternative approaches in the domain of psychopathology assessment. In creating a new mood measure, Baker and collegues (2000) demonstrated that Samejima's GRM had greater statistical fidelity to the underlying construct compared to a competing polytomous model (Master's Partial Credit Model). Note as well that, regardless of model, Embretson and Reise (2000) caution that any polytomous IRT modeling must be conducted on a heterogeneous sample because such models will only reasonably fit data that contains sufficient endorsement of each response category across all items. Based on both simulation and applied research studies, sample sizes in the range of 500 is considered necessary, and samples in the range of 1000 and greater are more ideal (e.g., Dodd, Koch, & De Ayala, 1989; Ping, Shuliang, Haijing, & Zhou, 2006, abstract; Reise & Yu, 1990).

The statistical construct of information is an additional and important consideration in interpreting the results or fit of Samejima's GRM analyses (and other IRT models). Information is an IRT derived statistic that quantifies measurement precision and is, therefore, akin to reliability estimates in CTT based analyses (Embretson & Reise, 2000). Measurement precision can be assessed at both the item and total test level through item and test information functions, respectively. Test information functions are "simply defined as the sum of the item information functions" (Dodd & De Ayala, 1994, p. 302). Information functions are an index of the standard error of measurement (SEM), as they are the reciprocal of the variance of the trait estimate (See results section for specific formulae). Given the direct relationship between SEM and information functions, Embretson and Reise emphasize that test information is a "critically important" statistic (p. 184). As reviewed by Dodd and De Ayala, in Samejima's GRM model, item and test information functions take into consideration both the discrimination power and the difficulty of an item/s. Conceptually, the more information a test/item yields, the more likely the item's difficulty exactly matches a respondent's true ability and demonstrates convergence with maximal discrimination (Embretson & Reise, 2000; Partchev, 2004; Samejima, 1977).

Limitations of IRT

Although IRT is practically useful and can overcome several shortcomings of CTT, IRT does not resolve or address all relevant measurement issues. Lord (1980) notes that various examinee and test characteristics can prevent IRT methods from effectively modeling the data. Specifically, examinee fatigue, illness, time pressure, and haphazard or otherwise random responding are detrimental to accurately fitting an IRT model. Also, IRT models may change with time. Bejar (1983) notes that despite the fact that IRT encompasses the principle of

"invariant item parameters" (p. 3), qualities of the construct may change with time and, hence, the adequacy of the fit of a given model may concomitantly change with time or situation. Bejar suggests that the fit of a given model must be constantly monitored over time. As repeatedly emphasized by Wright, use of non-Rasch based IRT models limits the degree to which truly objective measurement can be achieved (e.g., true sample and item independence will not be attained). Finally, in Embretson and Reise's (2000) experience, "raw scores and trait level estimates always correlated greater than .95" (p. 324). Hence, employing IRT methods will not necessarily produce markedly different measurement scales than typically achieved through CTT based approaches. Notwithstanding, the potential for IRT to greatly enhance measurement precision beyond that afforded through CTT and, in turn, inform diagnosis and future theory building remains. This is particularly important in the realm of PD measurement where advances in diagnostic precision, theory, and broad-band screening are sorely needed.

Present Study

Overview

Some CTT assumptions about the nature of measurement scales and their relation to hypothetical latent constructs are unrealistic. In particular, the proposition that each test item is equivalently related to the latent construct seems improbable (DeVellis, 1991). As a consequence, it is likely that measurement data obtained through assessment of a latent construct would either outright violate several CTT model assumptions or be less optimally captured by a scale derived within a CTT framework. An important consequence is that if estimates of the tool's reliability and validity are obtained within a CTT framework, but that framework does not aptly model the latent construct, the respective estimates may be artificially inflated or otherwise inaccurate.

Alternative means to conceptualize and evaluate the utility of measurement tools may prove more appropriate. IRT is one such alternative that has demonstrated substantial practical utility (e.g., Woodcock-Johnson III Tests of Cognitive Abilities; McGrew & Woodcock, 2001). Despite being readily available since approximately the early 1950s (e.g., Lord, 1953), the only large scale application of the theory has been in the realm of cognitive/intelligence and academic achievement testing. As highlighted by several authors (e.g., Embretson & Reise, 2000), psychopathology and personality assessment scholars have been very slow to adopt IRT methods. Given the potential wealth of psychometric information that can be gleaned through IRT compared to CTT methods, and the evidence to suggest that IRT is a "more theoretically justifiable" approach with a "greater potential to solve practical measurement problems" (Embretson & Reise, 2000, p. 3), it seems appropriate and well justified to utilize IRT based methods in the current test development exercise. Adhering to a rational-empirical strategy (e.g., Morey et al., 1985), PAI items will be conceptually classified into new scales according to *DSM-IV-TR* personality disorder criteria sets. From a psychometric standpoint, the appropriateness of each conceptually derived subscale will then be empirically evaluated through IRT methodology.

Method

Participants

A convenience sample of participants from clinical and nonclinical populations was solicited from archival databases of university student, outpatient, and inpatient populations who had previously completed PAI test protocols. The initial combined sample had 2435 participants (46% male, 48% female, 6% unknown). However, as recommended by Morey (1991), cases with significantly elevated scale scores on any of the four validity indexes (ICN, INF, NIM, PIM) were considered invalid and deleted from further analyses. Such cases are significantly more likely to indicate an inconsistent, haphazard, or intentionally distorted response style (e.g., faking good or bad). As a result, 307 invalid cases were omitted, and the sample size was reduced to 2128 participants. A listing of the individual data collection sites, and a description of the sample characteristics within each of the sites is detailed in Table 6. Approvals from Lakehead University's and, where applicable, the participating hospitals' respective ethics review boards were obtained.

Table 6

Sample Demographic Information

Site	Composition	Sample Size	Gender[a] (%)*
Lakehead University	Undergraduate students	871	231 (26.5) M 639 (73.4) F 1 (0.1)*
Correctional Facility	Inmate psychiatric referrals	413	368 (92.7) M 45 (7.3)*
St. Joseph's Hospital	Outpatient neuropsychological referrals	222	130 (58.6) M 92 (41.4) F
Lakehead Psychiatric Hospital	Combined in- and outpatient general psychiatric referrals	180	90 (50.0) M 89 (49.4) F 1 (0.6)*
Thunder Bay Regional Hospital	Combined in- and outpatient general psychiatric referrals	70	70 (100)*
Ottawa Hospital	General outpatient psychiatric referrals	372	149 (40.1) M 210 (56.5) F 13 (3.5)*
	Total Sample	2128	968 (45.5) M 1031 (48.4) F 129 (6.0)*

Note. [a]F = female, M = male, * = unknown: gender information is unavailable

Procedure

Following a procedure similar to that used by Morey, Waugh, and Blashfield (1985) in their construction of the PD scales for the MMPI, a two phase, rational-empirical strategy was employed in constructing the PD scales for the PAI. It was anticipated that both items that have only a subtle or more tangential association with a PD as well as items with a more obvious or strong association with a given PD would be encountered (e.g., Childs, Dahlstrom, Kemp, & Panter, 2000). Both types of items were rated at the rational stage of scale construction. The strength of each item's unique relation to the latent construct of the intended scale was later empirically evaluated. Retention of items in the final scales was ultimately determined by the results of the IRT based analyses.

Design: (a) Rational Component. Adhering to the ten PD criteria sets from the *DSM-IV-TR*, PAI items were conceptually rated (see above) into respective PD scales by several raters. Specifically, the prototypicality of each item for each PD was rated on a scale from –5 (= the exact opposite of the PD) through 0 (= irrelevant to the PD) to +5 (= highly prototypical of the PD). The raters were all psychology students ($N = 6$; four graduate and two undergraduate students). Level of expertise was somewhat variable across raters. For example, three raters were senior PhD students, had completed advanced courses in abnormal psychology, and had clinical experience working with individuals diagnosed with PD/s. Because the others were less knowledgeable, all raters received training: All raters were required to (a) review the PD section of the *DSM-IV-TR*; (b) re-read the *DSM-IV-TR* diagnostic criteria set for the respective PD immediately before rating the PAI items for that PD; and (c) complete a sample exercise that

involved rating the prototypicality for each PD for 25 items from a different personality test. After the sample exercise, each rater then attempted the same process with each of the actual PAI items: Raters estimated the prototypicality of each PAI item for one of the 10 PDs, and then repeated this process for a second PD until all 10 were completed. As a manipulation check, intraclass correlations were computed as an index of interrater reliability. As reviewed by Howell (1997), an intraclass correlation is akin to an effect size measure. It identifies the proportion of reliable variance (variance across item ratings) relative to the proportion of variance attributable to differences across judges and the interaction between judges and item ratings. Items with the highest (mean) prototypicality ratings across all raters for each PD were clustered into 10, preliminary PD scales.

Design: (b) Empirical Component. In the second phase, the conceptually derived, preliminary PD scales were refined through a series of empirically based analyses. Details of the proposed analyses are described in the subsequent analysis section. Briefly, the empirical process began with attempts to independently fit a graded response IRT model to each of the conceptually derived scales. This process yields several statistics that each provide useful information to address two primary research questions: (a) psychometric soundness of the individual test items, and (b) psychometric soundness of the total scales. No single index is definitive or superior to all others in this process. Strong items will discriminate well between respondents at different trait levels. Strong tests will consist of highly discriminating items that (a) provide coverage of the full trait range, and (b) assess a single underlying latent construct (single PD). Calibration of the respective scales started by inclusion of only the most prototypical items. Though an iterative process, additional items were added (based on prototypicality ratings) and poor performing items (e.g., provide limited information) were

removed. This iterative process continued until satisfactory scales were constructed for each of the 10 PDs, or all reasonable attempts were exhausted. For example, because the available item pool was not originally created to assess all 10 PDs, it was conceivable that less than 10 psychometrically sound scales would be created.

Analyses. Given that the PAI has a Likert response format, as Samejima's (1997) review suggests, a polytomous IRT model appears the most appropriate for the PAI dataset. The empirical process began with attempts to independently fit a graded response model to each of the 10 conceptually derived scales. Using the "ltm" module of the "R" statistical program (Rizopoulos, 2006a), Samejima's graded response model was used in the following form:

$$\log\left(\frac{\gamma_{ik}}{1-\gamma_{ik}}\right) = \beta ik + \beta iz \qquad (1)$$

"where γ_{ik} denotes the cumulative probability of a response in category kth or lower to the ith item, given the latent ability z". Akin to an item difficulty estimate, "βik's are the extremity" or threshold parameters (Rizopoulos, 2006a, p. 16; 2006b, p. 3). This calibration approach estimates theta (latent ability or, here, PD trait) using a Maximum Likelihood method and produces the following item and test statistics:

(a) Option Characteristic Curves (OCCs):

$$P_{ki}(\theta) = P^*_{ki}(\theta) - P^*_{(ki+1)}(\theta), \tag{2}$$

$$P^*_{ki}(\theta) = 1 / (1 + \exp(- D_{ai}(\theta - b_{ki+1}))),$$

where P = probability, k = response category, theta = trait, i = item, D = constant (equivalent to c or guessing parameter in 3PL), a = slope or discrimination parameter, and b = difficulty or response category threshold parameter.

(b) Item and Test Information Functions (IIF, TIF; 3.2 and 3.3, respectively), which are derived from item response information functions (3.1):

$$I_{xg}(\theta) = - \frac{\partial^2}{\partial\theta^2} \log P_{xg}(\theta), \tag{3.1}$$

$$I_g(\theta) = E[I_{xg}(\theta)] = \sum_{xg=0}^{mg} I_{xg}(\theta) P_{xg}(\theta), \tag{3.2}$$

$$I(\theta) = \sum_{g=1}^{n} I_g(\theta), \tag{3.3}$$

where $I_{xg}(\theta)$ indicates the amount of information in the test about the trait score that is provided by" a given response option to a single item; $I_g(\theta)$ "indicates the amount of information in the test about the trait score that is provided by a single item"; and $I(\theta)$ "indicates the amount of

information in the test about proficiency" or the trait score that is provided by all of the respective test items (Ramsay, 2000, p. 35; Samejima, 1977, p. 163).

As previously stated, no single parameter or statistic provides a definitive index of whether to retain or exclude a given item or scale. Consequently, item retention was decided based on concurrent consideration of all of the available indexes. As reviewed by Santor and Ramsay (1998), visual inspection of the OCCs provides a substantial amount of information: quality of discrimination over the full range of theta; thresholds for endorsing a given response option; and an indication of the ordinal property of each item's response options (or lack thereof). As well, as a quantitative index, discrimination parameters in the range of 1.0 or greater are generally accepted as strong, and strong scales will contain items with threshold parameters (theta intercept) that span the range of –2 to +2 SDs (e.g., Cooke & Michie, 1997; Rouse, Finger, & Butcher, 2000). Also, the item and test information functions provide the most direct index of the precision of measurement or reliability of individual items and the total test/s because the standard error of measurement is the square root of the reciprocal of the information functions. A search of the literature failed to identify an accepted quantitative standard to use as an approximate criterion to compare the obtained item and test information values. Rather, visual inspection of the item and test information functions (summation of item information functions) is the acceptable standard (e.g., Ramsay, 2000; Rizopoulos, 2006a; Toit, 2003).

As well, the assumptions of unidimensionality and local independence of the proposed PD scales were tested (Embretson & Reise, 2000). As reviewed and recommended by Meijer and Baneke (2004), unidimensionality and monotonicity will be assessed by computing scalability coefficients (*H*). Using MSP5 programming (Molenaar & Sijtsma, 2000), *H* coefficients (function of the sum of item pair covariances as a proportion of total scale covariance) were

computed to provide an index of discrimination strength across the test items. The *H* "coefficients are given by"

$$H_{gh} = \frac{Cov\,(X_g, X_h)}{Cov_{max}\,(X_g, X_h)}, \tag{4.0}$$

$$H_g = \frac{\sum_{h \neq g} Cov\,(X_g, X_h)}{\sum_{h \neq g} Cov_{max}\,(X_g, X_h)}, \tag{4.1}$$

$$H = \frac{\sum\sum_{g<h} Cov\,(X_g, X_h)}{\sum\sum_{g<h} Cov_{max}\,(X_g, X_h)} \tag{4.2}$$

where H_g and H are the scalability coefficients for a given item and the "total set of items in the test", respectively. And, "*Cov* denotes the covariance and X_g and X_h denote the item scores *g* and *h*, respectively" (p. 358). Essentially, the scalability coefficient is a "nonparametric analogue to the *a* parameter from logistic IRT models", is based on Guttman scaling principles (e.g., unidimensional, ranked), and is a "strictly increasing function of the variance of the total score" (Meijer & Baneke, p. 358). Based on Mokken's (1971) precedent, Meijer and Baneke suggest interpreting *H* coefficient values of .3 to .4 as a lower bound estimate for scale construction, values of > .4 and < .5 as moderate estimates, and values ≥ .5 as strong estimates of a unidimensional scale.

Lastly, to aid interpretation of the results in more familiar language, the traditional CTT based coefficient alpha estimate of internal consistency will also be computed for each of the final scales. It is generally accepted that internal consistency reliability estimates in the range of

.70 to .79 are the minimally acceptable or lower bound standard. Alphas in the range of .80 to .89 are considered moderate to strong. Alphas $\geq$ .90 are considered strong estimates and, more importantly, acceptable for diagnostic purposes (e.g., Flanagan et al., 2000; Kaplan & Saccuzzo, 1993). Note, however, it has also been argued that, in and of itself, a higher alpha estimate is not necessarily better. Rather, it is more important to interpret the alpha value within the context of additional qualitative and psychometric indicators (e.g., Cortina, 1993).

Results

Rational Phase

Inter-rater Reliability

All six raters completed the prototypicality ratings for each of the 10 respective PDs. Intraclass correlations were computed as an index of interrater reliability. Results are displayed in Table 7.

Table 7

Average Measure Intraclass Correlation Coefficients

PD	Average Intraclass Correlation Coefficient[a]
Paranoid	.87
Schizoid	.87
Schizotypal	.87
Antisocial	.91
Borderline	.91
Histrionic	.80
Narcissistic	.83
Avoidant	.87
Dependent	.88
Obsessive-Compulsive	.81

Note. [a]All correlations are significant at $p < .001$

As illustrated in Table 7, the average intraclass correlation coefficient estimates of interrater reliability are moderate/good to strong/excellent across all PDs. Because all estimates are significant, this suggests that a substantial proportion of reliable variance in the mean item ratings can be attributed to expected differences in item ratings as opposed to error variance (e.g.,

inconsistency within or across raters). It appears reasonable to use the mean of the judges' ratings as the prototypicality index in the empirical phase of the PD scale construction.

Empirical Phase

Sufficient items with high absolute value prototypicality ratings were identified to permit the creation of 10 new PD scales. With the exception of the HIS, NAR, and COM scales, sufficient items with a mean, absolute value prototypicality rating of 4.0 or greater were obtained for each PD scale (e.g., approx. 30-50 items per PD). Obtaining sufficient items for the HIS, NAR, and COM scales required using items with an absolute value rating of 3.0 or greater. For many of the scales, however, substantial item overlap was encountered. In particular, item ratings for all three of the Cluster A PDs (PAR, SZD, SZT) and the AVD (Cluster C/ internalizing PD) had substantial overlap (see Appendix C). As a consequence, an additional design manipulation was employed.

For the university student participant sample (n = 871, 26.5% male), personality assessment ratings were also available for a second personality measure: The Millon Clinical Multiaxial Inventory (third edition; MCMI-III; Millon, Davis, & Millon, 1997). The MCMI-III is a 175 item, broadband psychopathology inventory with a true-false response format, and 28 subscales. Ten subscales assess each of the 10 *DSM-IV-TR* PDs. An additional four subscales assess PDs that are more consistent with Millon's own theory of personality disorders. The MCMI-III also contains other clinically related scales, but only the *DSM-IV-TR* PD scales were examined here. Specifically, in addition to high prototypicality ratings, correlations between each PAI item and the respective MCMI-III PD scale were also taken into consideration in creating

the new, PAI PD scales. High prototypicality ratings always took precedent and then, if needed, scales were augmented with items that had high correlations with the MCMI-III PD scales.

For example, on the preliminary AVD scale, only seven of 24 items with high prototypicality ratings (≥ 4.0) did not overlap with PAR, AVD, or DEP scales. Consequently, when running the IRT analyses, the conceptually derived scale was augmented with PAI items that had high correlations with the MCMI-III AVD scale. Retention of items on the scale remained determined by results of the IRT based analyses. Note as well, however, given that substantial item overlap was encountered, if an item fit quite well (from an IRT standpoint) on more than one scale, a qualitative assessment of the respective item was again taken into consideration. Overall, the PD scale construction process was similar to that described by Morey (1991) in creation of the original PAI scales: A rigid, rational-empirical paradigm did not prove useful. After the initial conceptual ranking procedure, the item inclusion and retention process became very iterative. Several psychometric indexes were taken into consideration, and no single index was definitive. Resolving item overlap became the most challenging obstacle. A detailed description of the results for each respective scale is detailed below[1].

[1] Note: The statistical methods employed in this project generated substantial amounts of data. Given that the result section involves reviewing findings for over 10 scales, a decision was made to incorporate the scale specific discussion within the results section. A more general discussion section subsequently follows, which includes a discussion of the more overarching themes, conclusions, and implications of this scale development process as a whole.

PAR

Items for the final version of the PAR PD scale are displayed in Table 8, and the original PAI subscale membership for each item is also noted.

Table 8

PAR Scale Items

Original PAI Subscale	PAR Item description (PAI item number)
PAR-P	(269) People have had it in for me.
PAR-P	(69) Some people do things to make me look bad.
PAR-P	(189) There are people who want to hurt me.
PAR-P	(149) Some people try to keep me from getting ahead.
PAR-R	(157) People don't appreciate what I've done for them.
PAR-P	(29) Certain people go out of their way to bother me.
PAR-H	(208) People think I'm too suspicious.
PAR-H	(168) People generally hide their real motives.
ANX-A	(204) I often feel as if something terrible is about to happen.
BOR-N	(179) I've made some real mistakes in the people I've picked as friends.
PAR-P	(109) People around me are faithful to me.
BOR-N	(99) People once close to me have let me down.
PAR-H	(48) I have been alert to the possibility that people will be unfaithful.
PAR-H	(8) Most of the people I know can be trusted.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the composite indexes or initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .87; total scale *H* coefficient was .39; and the individual item *H* values (H_g) ranged from .36 to .45. The alpha estimate is moderate to strong, particularly given that the scale has 14 items, and supports that the PAR scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable, and falls in the low to moderate range. Inspection of the individual H_g values revealed that all were acceptable and six of the 14 items fell in the moderate range. When interpreted in context with the IRT results that follow, these findings support that the PAR scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the graded response model (GRM) was obtained for the PAR items. Results of the IRT analyses are displayed graphically in Figures 5 through 7, and statistically in Tables 8.1, 8.2, and 8.3 (see Appendix D). The CCCs illustrate the probability of endorsement of a given response category as a function of the PAR PD trait. Ideal CCCs will have narrow and peaked curves that span the full range of trait levels. Item 168 is an example of a strong item: The valence of item endorsement corresponds near precisely with increasing levels of the PAR trait, and four distinct curves are easily discernable. Items 69, 189, and 269, however, ultimately contribute more information because their discrimination parameters are much stronger (see Table 8.1). Of interest, these three items almost function as dichotomies. A very steep α parameter is noticeable in the CCCs for both the total item endorsement and first category thresholds. As well, the intercepts for the first category thresholds fall in the trait range of 0.0 to 0.60. This suggests that even moderate endorsement of these items is indicative of higher levels of PAR trait. This finding is further supported in the item and test information functions. Overall, the PAR scale items have acceptable CCCs: Within each item, the category thresholds

demonstrate an ordinal property and fall within acceptable theta ranges (M = -.33, 1.0, and 2.04 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.51, *range* = 1.01 to 2.13).

The item and test information functions corroborate that the PAR items assess a broad range of the PAR trait levels. Some items are clearly more informative than others (e.g., 269, 69, 189), however, all contribute reliable information. The individually, less informative items are necessary because they provide information at the low to mid range of theta. The higher information items assess the mid to upper range of theta. The PAR scale provides more information and, hence, finer discrimination across respondents at the higher range of theta (e.g., 62% within theta range of 0 to +3, vs. 27% within the theta range of -3 to 0). Although a more balanced representation of items across the full range is desirable, additional PAI items that assessed the same PAR trait, but at lower trait levels could not be identified.

In sum, the total scale indexes for the PAR scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. As illustrated in Table 8, most items that fit on the PAR PD scales are from the Paranoia Scale of the original PAI. Of interest, the items spanned more than one subscale and included items from the original Borderline Features and Anxiety scales. This suggests that the new scale is not simply replicating the original Paranoia scale, which provides support that the new scale assesses a distinct trait. Evidence of content validity is strong: All items are clearly consistent with the *DSM-IV-TR* criteria set for PAR PD. The scale appears to assess several core features of the *DSM* criteria set: pervasive distrust, undue suspiciousness, and biased assumption of malevolent motivations. Less well represented are two remaining *DSM* features: tendency to bear grudges and hypersensitivity to character attack. Overall, the PAR scale demonstrates strongest psychometric properties at the

mid to higher end of the trait continuum and should, therefore, provide accurate clinical screening and delineation of functioning for individuals within the clinical range. Scale scores for individuals below minus one standard deviation will be less reliable. Consideration of both content validity and the quantitative findings suggests that the PAR scale could be used for applied research and clinical purposes.

SZD

Items for the final version of the SZD PD scale are displayed in Table 9, and the original PAI subscale membership for each item is also noted.

Table 9

SZD Scale Items

Original PAI Subscale	**SZD Item description (PAI item number)**
SCZ-S	(270) I make friends easily.
WRM	(13) I'm a very sociable person.
SCZ-S	(190) I enjoy the company of other people.
WRM	(53) It's easy for me to make new friends.
WRM	(93) I like to meet new people.
SCZ-S	(230) I like to be around other people if I can.
SCZ-S	(110) I'm a loner.
SCZ-S	(30) I just don't seem to relate to people very well.
SCZ-S	(310) I keep in touch with my friends.
NON	(81) If I'm having problems, I have people I can talk to.
SCZ-S	(70) I don't have much to say to anyone.
WRM	(333) I have more friends than most people I know.
NON	(121) I spend most of my time alone.
RXR	(202) I'm comfortable with myself the way I am.
SCZ-S	(150) I don't feel close to anyone.
BOR-N	(139) I rarely feel very lonely.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .92; total scale *H* coefficient was .46; and the individual item *H* values (H_g) ranged from .38 to .53. The alpha estimate is very strong and supports that the SZD scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is also acceptable and falls in the moderate range. Inspection of the individual H_g values revealed that all but one item fell in the moderate range, and three items fell in the optimal/strong range (pr270, pr13, and pr190). Considered in context with the IRT findings that follow, these results support that the SZD scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the SZD items. Results of the IRT analyses are displayed graphically in Figures 8 through 10, and statistically in Tables 9.1, 9.2, and 9.3 (see Appendix E). Inspection of the IRT results indicate that the SZD scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -.69, 0.61, and 1.75 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.66, *range* = 1.04 to 2.64). The item and test information functions corroborate that the SZD items assess a broad range of the SZD trait levels. The results are strong as the full range of theta is well represented with highly discriminating items. It is noted that the IRF curves for the items with the highest information (e.g., pr270, pr13) are not smooth. Inspection of the individual item response CCCs suggests that, rather than problematic, the nonsmooth IRF results are directly attributable to the highly discriminating quality of these items (see Appendix E). If the threshold for endorsement of each item response category is very strong (steep, highly discriminating CCCs), the overall function becomes saw-toothed as opposed to smooth.

Overall, the SZD scale provides more information and, hence, finer discrimination across respondents at the higher range of theta (e.g., 57% within theta range of 0 to +3, vs. 35% within the theta range of -3 to 0). However, the information is comparable within one standard deviation above and below the mean (23% and 20%, respectively). Thus, the reliability evidence for the SZD scale is strong for individuals across a wide range of SZD trait. And, as a clinical tool, the SZD scale will likely discriminate well among individuals at the highest range of theta (e.g., > 1 *SD*). There is a slight under representation of highly discriminating items in the theta range of less than 1.5 standard deviations below the mean. With respect to clinical utility, given that the scaling limitation is mild and impacts the assessment of individuals at the extreme low-end of the trait distribution, this limitation is not markedly significant.

In sum, the total scale indexes for the SZD scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. As illustrated in Table 9, many items that fit on the SZD PD scale are from the Schizophrenia (Social Detachment Subscale) and the Interpersonal (Warmth) scales from the original PAI. Evidence of content validity is strong: All items are clearly consistent with the *DSM-IV-TR* criteria set for SZD PD. Each item relates to some form of low social need or poor social facility. Six items are the most informative and suggest, consistent with SZD PD, that difficulty making friends and lack of enjoyment from social contact are the hallmark features of the SZD scale. The second core feature of SZD PD, restricted emotionality/aloofness, is less sufficiently captured. Overall, consideration of both content validity and the quantitative findings supports that the SZD scale can be used for both applied research and clinical purposes.

SZT

Items for the final version of the SZT PD scale are displayed in Table 10, and the original PAI subscale membership for each item is also noted.

Table 10

SZT Scale Items

Original PAI Subscale	**SZT Item description (PAI item number)**
BOR-I	(57) Sometimes I feel terribly empty inside.
DEP-C	(67) Sometimes I think I'm worthless.
ANX-C	(65) It's often hard for me to enjoy myself because I am worrying about things.
DEP-A	(46) I've forgotten what it's like to feel happy.
ANX-A	(4) I am so tense in some situations that I have great difficulty getting by.
ANX-C	(105) I'm often so worried and nervous that I can barely stand it.
DEP-A	(126) Nothing seems to give me much pleasure.
SCZ-T	(38) My thinking has become confused.
ANX-C	(27) I feel that I've let everyone down.
DEP-C	(25) I often have trouble concentrating because I'm nervous.
DEP-C	(107) I don't feel like trying anymore.
DEP-C	(187) No matter what I do, nothing works.
DEP-A	(86) Everything seems like a big effort.
ANX-A	(44) I can't do some things well because of nervousness.
NIM	(169) People don't understand how much I suffer.
RXR	(2) I have some inner struggles that cause problems for me.
SCZ-T	(118) Sometimes I have trouble keeping different thoughts separate.
ARD-O	(45) I have impulses that I fight to keep under control.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the composite indexes or tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .94; total scale *H* coefficient was .51; and the individual item *H* values (H_g) ranged from .47 to .57. The alpha estimate is exceptionally strong, to the extent that singularity might be questioned. Inspection of the individual inter-item correlations, however, reveals that no correlations were greater than .70, and most correlations fell in the range of .37 to .60. Consistent with the high alpha estimate, the total scale *H* coefficient is also strong. Inspection of the individual H_g values revealed that all items were acceptable. All fell in at least the moderate range, and most were strong (e.g., all H_g values were $\geq$ .47). When interpreted in context with the IRT results that follow, these findings support that the SZT scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the SZT items. Results of the IRT analyses are displayed graphically in Figures 11 through 13, and statistically in Tables 10.1, 10.2, and 10.3 (see Appendix F). Inspection of the IRT findings indicate that the resulting SZT scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (*M* = -.40, 0.70 and 1.63 at β1, β2, and β3, respectively). As well, all α parameters are strong (*M* = 1.79, *range* = 1.41 to 2.33). The item and test information functions corroborate that the SZT items are highly discriminating across a broad range of the SZT trait levels: A total of 80% of the test information falls within two standard deviations above and below the mean. Overall, the SZT scale provides more information and, hence, more precise discrimination across respondents through the midrange of theta (e.g., 49% within theta range of 0 to +2, and 32% within the theta range of -2 to 0). Notwithstanding, as a clinical tool, the SZT scale will also likely discriminate well among individuals at the highest range of theta (e.g., 35% of the total test information falls $\geq$ 1 *SD*).

Overall, like the SZD results, there is evidence of a mild under representation of highly discriminating items in the theta range of less than 1.5 standard deviations below the mean. With respect to clinical utility, however, given that the scaling limitation is mild and impacts the assessment of individuals at the extreme low-end of the trait distribution, this limitation is not markedly significant. In sum, the quantitative indexes for the SZT scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure.

As illustrated in Table 10, items that fit on the SZT PD scale are from a host of scales on the original PAI. Evidence of content validity with respect to assessing a single dimension is strong: The items appear to assess a single dimension of internal distress related to thoughts and feelings of apathy and worthlessness (withdrawal, loneliness, sadness, anxiety). Several items were drawn from the Depression – Cognitive subscale. Despite that ideas of reference, magical thinking, and altered perception are hallmark SZT PD features, only two items from the original SCZ scale and no PAR items fit on this scale. This finding is primarily attributable to the decision to eliminate item overlap as much as possible across the new PD scales. Many original SCZ and PAR items fit on each of the SZD, SZT, and PAR PD scales in the original IRT runs. Through the iterative process previously described, the SCZ and PAR items were primarily assigned to the SZD and PAR PD scales, respectively. Second, IRT analyses of the original PAR and SCZ scales revealed that disordered thought process related items appear to tap a relatively discrete, unidimensional phenomenon. In particular, the items do not share marked variance with social detachment. Thus, it is probable that more PAR and SCZ items from the original PAI did not model with the SZT PD scale here because the remaining items more likely assess a specific psychotic spectrum or related thought disordered process.

AVD

Items for the final version of the AVD PD scale are displayed in Table 11, and the original PAI subscale membership for each item is also noted.

Table 11

AVD Scale Items

Original PAI Subscale	AVD Item description (PAI item number)
ANX-C	(265) I usually worry about things more than I should.
PIM	(24) Sometimes I let little things bother me too much.
DEP-A	(6) Much of the time I'm sad for no real reason.
ARD-P	(26) I often fear I might slip up and say something wrong.
ARD-P	(66) I have exaggerated fears.
BOR-A	(94) My mood is very steady.
PIM	(184) I don't take criticism very well.
BOR-A	(174) I've always been a pretty happy person.
ARD-P	(106) I get very nervous when I have to do something in front of others.
ANX-A	(244) I seldom feel anxious or tense.
DOM	(216) I prefer to let others make decisions.
WRM	(173) It takes me a while to warm up to people.
WRM	(213) It takes a while for people to get to know me.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .83; total scale *H* coefficient was .35; and the individual item *H* values (H_g) ranged from .29 to .40. The alpha estimate is moderate and supports that the AVD scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is

acceptable, and falls in the low range. Inspection of the individual H_g values revealed that all but one item (pr244, H_g = .29) demonstrated acceptable values, and the remaining scale items fell within the lower bound range. When interpreted in context with the IRT results that follow, these findings support that the AVD scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the AVD items. Results of the IRT analyses are displayed graphically in Figures 14 through 16, and statistically in Tables 11.1, 11.2, and 11.3 (see Appendix G). Inspection of the IRT results indicates that the AVD scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -1.49, 0.09, and 1.57 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.24, *range* = 0.87 to 1.67). The item and test information functions corroborate that the AVD items discriminate well across the full trait range. The percent of test information is near equally distributed above and below the mean (45% within theta range of 0 to +3, and 41% within -3 to 0). Measurement precision is, therefore, strong across both low and high trait levels.

As illustrated in Table 11, items that fit on the AVD PD scales were drawn from several original PAI scales. Evidence of content validity is strong with respect to assessing a single internalizing dimension. The items reflect an internalizing dimension that seems to measure a general tendency to experience fear, worry, and avoidance. Items reflecting other core features of AVD PD (e.g., fear of interpersonal contact re evaluation, shame/inadequacy, ridicule) are less sufficiently captured. Overall, consideration of both the qualitative and quantitative findings suggests that the total scale indexes for the AVD scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. The results suggest that

both low functioning individuals on this trait and those within the clinical range will be reliably assessed.

ANT

Three strategies were used in creating and analyzing the ANT scale data. First, IRT analyses were run on the original ANT total and subscale structure as defined in the manual. Second, because psychometric limitations were identified in this first process, the original items were reconfigured based on the results of the IRT analyses. The aim of this strategy was to devise a new ANT scale with stronger psychometric properties that still contained only the original ANT items. Third, an attempt was also made to create an entirely new ANT scale using the same rational-empirical process employed for the other PD scales. The goal of the third approach was to explore whether the original ANT scales could be improved through augmentation with other items. Results of the three strategies follows.

Original ANT Scales. IRT analyses were run on both the original PAI ANT total scale and the respective subscale structure. Items for the original ANT total scale arranged in descending order based on the amount of information contributed to the total scale are displayed in Table 12. The original PAI subscale membership for each item is also noted.

Table 12

ANT Scale Items

Original PAI Subscale	ANT Item description (PAI item number)
ANT-S	(79) I do a lot of wild things just for the thrill of it.
ANT-S	(119) My Behaviour is pretty wild at times.
ANT-S	(39) I get a kick out of doing dangerous things.
ANT-A	(171) I like to see how much I can get away with.
ANT-E	(71) I'll take advantage of others if they leave themselves open to it.
ANT-A	(51) I've deliberately damaged someone's property.
ANT-A	(131) I used to lie a lot to get out of tight situations.
ANT-A	(91) I've done some things that weren't exactly legal.
ANT-S	(279) I'm not a person who turns down a dare.
ANT-E	(111) I'll do most things if the price is right.
ANT-E	(151) I can talk my way out of just about anything.
ANT-E	(31) I've borrowed money knowing I wouldn't pay it back.
ANT-S	(239) I like to drive fast.
ANT-A	(291) I've never taken money or property that wasn't mine.
ANT-S	(159) If I get tired of a place, I just pick up and leave.
ANT-S	(319) I never take risks if I can avoid it.
ANT-E	(311) When I make a promise, I really don't need to keep it.
ANT-A	(11) I was usually well-behaved at school.
ANT-E	(231) I don't like to stay in a relationship very long.
ANT-S	(199) The idea of "settling down" has never appealed to me.
ANT-A	(211) I was never expelled or suspended from school when I was young.
ANT-A	(251) I've never been in trouble with the law.
ANT-E	(271) I look after myself first; let others take care of themselves.
ANT-E	(191) I don't like being tied to one person.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .87; total scale *H* coefficient was .31; and the individual item *H* values (H_g) ranged from .22 to .39. The alpha estimate is moderate to strong and supports that the ANT scale items demonstrate acceptable inter-item correlations. The total scalability coefficient falls at the lower bound criterion and is, therefore, minimally acceptable. Inspection of the individual H_g values revealed that 13 of 24 items fell below the acceptable range and the remaining items fell in the lower bound range. These findings are consistent with problems identified in the IRT results that follow. The incongruence between the strong alpha value and weak individual and total *H* values fails to support the unidimensional assumption and suggests that the original ANT scale is multifaceted. This finding is not entirely unexpected given that the ANT scale was specifically constructed to comprise three subtests. The low individual item H_g values and the problematic CCCs (see below), however, suggest that when IRT analyses are run on the individual subtests, some problematic psychometric properties may remain.

A reasonable fit of the GRM was obtained for some, but not all of the ANT items. Results of the IRT analyses are displayed graphically in Figures 17 through 19, and statistically in Tables 12.1, 12.2, and 12.3 (see Appendix H.1). Inspection of the IRT results indicates that the majority of the original ANT scale items have problematic item response CCCs: The desired ordinal property of the category thresholds is not always readily discernable (e.g., pr11, p191, p231, p271, p311; see Appendix H.1) and similarly, on average, the respective category thresholds fall outside acceptable theta ranges (*M* = -0.02, 1.23, and 2.51 at β1, β2, and β3, respectively). As well, many α parameters are problematic (*M* = 1.20, *range* = 0.53 to 2.50). Inspection of the individual item response CCCs suggests this is attributable to the presence of (a) items that are simply problematic overall (see examples listed above) and (b) items that

appear to function in a more dichotomous versus polytomous format (e.g., p51, p91, pr211, pr251, pr291; see Appendix H.1). The item and test information functions corroborate that the ANT total scale demonstrates problematic psychometric properties: Nine of the 24 items contribute the majority of the test information, and the test information function is negatively skewed (24% of the test information falls in the theta range of –3 to 0 vs. 58% in the theta range of 0 to +3).

Findings across the individual subscale analyses were mixed. Results of the initial tests of monotonicity and dimensionality for the ANT-A subscale were as follows: Cronbach's coefficient alpha was .80; total scale *H* coefficient was .42; and the individual item *H* values (H_g) ranged from .39 to .47. The alpha estimate is moderate (but impressive because it is derived from eight items) and supports that the ANT-A subscale items demonstrate acceptable inter-item correlations. The total scalability coefficient for the ANT-A falls in the moderate range. Inspection of the individual H_g values revealed that all items were acceptable. Four fell in the lower bound range and the remaining four fell in the moderate range. Overall, the total and individual item scalability results for the ANT-A subscale are demonstrably improved from the full, ANT scale results. These findings support that the ANT-A scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the ANT-A subscale. Results of the IRT analyses are displayed graphically in Figures 20 through 23, and statistically in Tables 12.4 through 12.7 (see Appendix H.2). Results for the ANT-A subscale were strong, particularly given that the scale has only eight items. The overall test information function was smooth and reasonably normally distributed, but remained shifted slightly to the higher end of theta. Specifically, evidence of negative skew was discernable as 60% of the total information fell in

the upper theta range (0 to +3), compared to 33% in the lower theta range (-3 to 0). Overall, the test information function analyses indicate that the ANT-A subscale provides more precise measurement across the full trait range compared to the total scale results (93% vs. 82% across theta range of -3 to +3, respectively).

The individual item response CCCs were acceptable. Four items were modestly improved over the total scale results (pr11, p91, pr251, pr291) and one item was slightly weaker (p171). Of interest, items pr211 and pr251 demonstrated near perfect dichotomous compared to polytomous response properties. Inspection of the response category frequency data revealed that this is attributable to the true/false nature of these specific items: "I was never expelled or suspended from school when I was young", and "I've never been in trouble with the law". Less than 16% of the respondents endorsed either of the middle two response options for these items. With the exception of these two items, the individual item category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = 0.01, 0.71, and 1.43 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.61, *range* = 1.07 to 2.25).
As a clinical tool, the ANT-A scale should adequately assess individuals across a reasonably wide-range of trait functioning. The measure will likely be particularly strong in assessing individuals at the high end of theta and, therefore, very useful with clinical populations. Alternatively, the scale will be less reliable in assessing individuals at the lower end of theta (< 1 *SD*).

Findings for the ANT-E subscale were less positive. Results of the additional tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .66; total scale H coefficient was .31; and the individual item H values (H_g) ranged from .29 to .37. The alpha estimate is low. Inspection of the inter-item correlations corroborates poor internal

consistency. Inter-item correlations range from .11 to .36. The total scalability coefficient for the ANT-E falls in the lower bound range. Inspection of the individual H_g values revealed that six of the eight items were acceptable, and all of the acceptable items fell in the lower bound range. Overall, the total and individual item scalability results for the ANT-E subscale are not markedly improved from the full ANT scale results. These findings suggest that the ANT-E scale demonstrates low monotonicity and unidimensional properties, which appears to be the result of the ANT-E items tapping almost exclusively the high end of theta (see IRT results). Moreover, the assessment of even the high theta functioning is less reliable than ideal.

Results of the IRT analyses for the ANT-E subscale were also problematic. The overall test information function was smooth, but markedly shifted to the high end of theta. Evidence of negative skew was readily discernable as 54% of the total information fell in the upper theta range (0 to +3), compared to 18% in the lower theta range (-3 to 0). The negative skew detrimentally impacts the reliability of the overall scale: The test information function indicates that the ANT-E subscale provides less precise measurement across the full trait range compared to the ANT total scale results (72% vs. 82% across theta range of -3 to +3, respectively). Similarly, the individual item response CCCs were problematic. Four subscale items were modestly improved over the total scale results (p191, p231, p271, p311), and the other four were essentially unchanged. The individual item category thresholds demonstrate an ordinal property, but cluster within the very high theta ranges (M = 0.27, 1.61, and 2.94 at β1, β2, and β3, respectively). In particular, the intercept for the first category threshold for four of the subscale items fell near one standard deviation above the mean (0.91 to 1.40 *SD*). All α parameters are moderate to strong (M = 1.17, *range* = 0.85 to 1.72). If these data are assumed to be valid, the ANT-E items appear to assess only a very extreme, pathological component of the ANT-E trait

spectrum. As a clinical tool, these results suggest that the ANT-E scale will only provide minimally acceptable, reliable assessment data for individuals at the high end of theta. Consequently, the ANT-E scale likely has low utility across both clinical and nonclinical populations.

The results for the ANT-S subscale were somewhat unusual. Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .77; total scale *H* coefficient was .36; and the individual item *H* values (H_g) ranged from .30 to .47. The alpha estimate is low, but may be reasonable for a screening tool given that the scale has only eight items. Inspection of the inter-item correlations (*range* = .06 to .69) suggests that although some correlations were moderate to strong (> .30), the low estimate of internal consistency is due to the presence of several weak inter-item correlations (< .20) on a brief scale. The total scalability coefficient for the ANT-S falls in the lower bound range, but is improved from the total scale *H* value. Inspection of the individual H_g values revealed that all items were acceptable. Five of eight fell in the lower bound range (three exactly at the minimally acceptable criterion), and the remaining three items fell in the moderate range.

Overall, the total and individual item scalability results for the ANT-S subscale are modestly improved from the full ANT scale results. These findings suggest that the ANT-S scale demonstrates minimally acceptable monotonicity and unidimensional properties. Given that this measure has only eight items and is intended to assess a very circumscribed domain of functioning, higher inter-item correlations would be expected. It is clear from the results that these eight items do not strongly represent a single dimension. Most likely, the total subscale indexes fell within acceptable limits because of the inclusion of three exceptional items. Given that the remaining majority of subscale items modestly correlate with these strong items but not

each other, it is unclear what construct the total subscale scores truly assess. It is conceivable that the five remaining items are more strongly pulling other domains of functioning. Further insights are provided by the IRT results.

A weak fit of the GRM was obtained for the ANT-S subscale. Inspection of the test and item information functions reveal the odd pattern of psychometric findings demonstrated by this subscale. The overall test information function is saw-toothed, steep (kurtotic), and shifted toward the high end of theta (mild negative skew). This is attributable to three, highly discriminating items (p79, p39, p119) that contribute an exceptional share (71%) of the reliable variance of the total scale. Inspection of the individual item CCCs reveals that the psychometric properties of these three items are very strong: The CCCs for each response option are relatively normally distributed and the ordinal response format is easily discernable. The category threshold discrimination parameters are very sharp (αs at ≥ 2.0). The strong discrimination, however, occurs at the cost of trait representation. Although each of these items are highly reliable, the measurement precision is only informative for a restricted range of moderate to high trait functioning (approximately -0.5 to +1.5 *SD*). This would be less problematic if the remaining items contributed reliable information in the neglected theta ranges.

Results of the IRT and nonparametric findings suggest three additional subscale items demonstrate acceptable psychometric properties (p239, p279, pr319). Because the three strongest items are exceptional, however, the contribution of the remaining scale items to the overall information function is somewhat masked or misrepresented in the graphic presentation (IIFs). Consideration of the individual item IRT parameters indicates these additional three items demonstrate reasonable ordinal properties and contribute some reliable variance, albeit low. As illustrated in the CCC graphs, the remaining two ANT-S items (p159, p199) are clearly

problematic across all statistical indexes. Inspection of the inter-item correlation matrix further supports this interpretation as items 159 and 199 demonstrate weak correlations with each other ($r = .23$), with the three high information items ($r \leq .31$) and, in particular, with the low/moderate information items ($r \leq .18$). Overall, despite demonstrating many acceptable quantitative properties, when all of the available indexes are considered, it appears that the ANT-S scale likely has low practical utility across both clinical and nonclinical populations.

Lastly, a qualitative review of the IRT results for the original ANT total and respective subscale analyses suggests that the total scale primarily assesses fairly extreme thrill-seeking behaviour, law-breaking behaviour, and self serving, opportunistic behaviour. The ANT-A scale appears to assess law and rule-breaking behaviour. The ANT-E scale appears to assess opportunistic, self serving, exploitative motivations or behaviours, and/or endorsement of antisocial values. The ANT-S scale appears to assess a very circumscribed aspect of thrill-seeking behaviour. In sum, however, it is difficult to interpret the clinical significance of these findings because, with the exception of the ANT-A scale, evidence of unidimensionality was less than ideal. Although some very sound items and evidence of a dominant single dimension were evident for each scale, a majority of items demonstrated either weak internal psychometric properties (e.g., poor CCCs) and/or unacceptably low evidence of cohesion with the dominant trait of interest.

Modified Original ANT Scale. In an attempt to better ascertain the primary dimension underlying the ANT total scale, IRT analyses were subsequently run on all of the ANT items. Problematic items were eliminated with a view to discerning a more pure, unidimensional trait component for the original ANT test items. Items for the final version of the modified original

ANT PD scale are displayed in Table 13, and the original PAI subscale membership for each item is also noted. A reasonable fit of the GRM was obtained for the ANT items. Results of the IRT analyses are displayed graphically in Figures 24 through 26, and statistically in Tables 13.1, 13.2, and 13.3 (see Appendix H.3).

Table 13

Original ANT Total Scale with Low Information Items Removed

Original PAI Subscale	**ANT Item description (PAI item number)**
ANT-S	(79) I do a lot of wild things just for the thrill of it.
ANT-S	(119) My Behaviour is pretty wild at times.
ANT-S	(39) I get a kick out of doing dangerous things.
ANT-A	(171) I like to see how much I can get away with.
ANT-A	(51) I've deliberately damaged someone's property.
ANT-E	(71) I'll take advantage of others if they leave themselves open to it.
ANT-A	(131) I used to lie a lot to get out of tight situations.
ANT-A	(91) I've done some things that weren't exactly legal.
ANT-E	(31) I've borrowed money knowing I wouldn't pay it back.
ANT-S	(239) I like to drive fast.
ANT-A	(291) I've never taken money or property that wasn't mine.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

As illustrated in Table 13, items that fit on the ANT PD scale were drawn primarily from the ANT-S and ANT-A subscales of the original PAI. One item is from the ANT-E scale. Removal of low information items and reanalyses with the IRT methods appears to have improved the original scale. The remaining items are near identical to the high information items

from the original IRT investigation of the total scale. Evidence of content validity is strong. The items appear to capture the essence of ANT: "...pattern of disregard for and violation of the rights of others" (*DSM-IV-TR*, 2000, p. 685) and, except for physical aggression, reflect the full ANT PD criteria set.

Inspection of the IRT results indicates that the ANT scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable, but somewhat high theta ranges (M = 0.05, 0.91, and 1.77 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.66, *range* = 1.09 to 2.53). The item and test information functions corroborate that the ANT items discriminate well across the mid to high range of the ANT trait. The mild negative skew or shifting of the test information function to the higher range of theta is evident as 67% of the test information falls in the upper theta range (0 to +3) whereas, 24% falls in the lower theta range (-3 to 0). Nonetheless, this briefer scale provides more precise information across the full trait range compared to the lengthier, original ANT total scale (91% vs. 82% of the total information falls within the theta range of -3 to +3, respectively).

Results of the additional tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .85; total scale H coefficient was .43; and the individual item H values (H_g) ranged from .33 to .47. The alpha estimate is moderate, and supports that the ANT scale items demonstrate acceptable inter-item correlations. The total scale H coefficient is acceptable and also falls in the moderate range. Inspection of the individual H_g values revealed that all but one item fell in the moderate range. When interpreted in context with the additional indexes reviewed, these findings support that the modified ANT scale demonstrates acceptable monotonicity and unidimensional properties.

With respect to clinical utility, the drop in measurement precision at the lower trait range (<1 *SD*) is likely not markedly significant because the measurement precision of individuals in the clinical range is acceptable. Lastly, of particular significance, this exercise demonstrates a notable strength of IRT methods: Because the IRT methods permit the specific psychometric properties of individual items to be identified, briefer but equally reliable scales can be devised to assess a given construct. In sum, this modified ANT scale derived only from original ANT items is briefer, yet appears to provide at least equivalent assessment of the latent trait. Further, measurement precision may actually be stronger with the briefer scale.

New ANT scale. The last strategy undertaken within this domain was an attempt to create an entirely new ANT scale using the same rational-empirical process employed for the other PD scales. The intent was to explore whether the original ANT scales could be improved through augmentation with other items. Items for this last version of the ANT PD scale are displayed in Table 14, and the original PAI subscale membership for each item is also noted.

Table 14

New ANT Scale Items

Original PAI Subscale	ANT Item description (PAI item number)
AGG-P	(101) Sometimes I'm very violent.
AGG-P	(181) I've threatened to hurt people.
AGG-A	(258) I have a bad temper.
AGG-P	(61) Sometimes my temper explodes and I completely lose control.
AGG-P	(21) People are afraid of my temper.
ANT-S	(119) My Behaviour is pretty wild at times.
BOR-A	(134) I have little control over my anger.
AGG-P	(141) Sometimes I smash things when I'm upset.
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.
BOR-S	(303) I'm a reckless person.
AGG-A	(338) When I get mad, it's hard for me to calm down.
BOR-A	(214) I've had times when I was so mad I couldn't do enough to express all my anger.
BOR-S	(223) I'm too impulsive for my own good.
ANT-A	(171) I like to see how much I can get away with.
ANT-A	(51) I've deliberately damaged someone's property.
BOR-A	(54) My moods get quite intense.
ANT-A	(131) I used to lie a lot to get out of tight situations.
ANT-S	(79) I do a lot of wild things just for the thrill of it.
ANT-S	(39) I get a kick out of doing dangerous things.
DRG	(182) I've used prescription drugs to get high.
BOR-N	(19) My relationships have been stormy.
DRG	(62) People have told me that I have a drug problem.
DRG	(22) Sometimes I use drugs to feel better.
DRG	(23) I've tried just about every type of drug.
ANT-A	(91) I've done some things that weren't exactly legal.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .93; total scale *H* coefficient was .43; and the individual item *H* values (H_g) ranged from .38 to .49. The alpha estimate is strong, and supports that the new ANT scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable and falls in the moderate range. Inspection of the individual H_g values revealed that five of the 22 items fall in the lower bound range, and the remainder fall in the moderate range. These findings support that the new ANT scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the new ANT items. Results of the IRT analyses are displayed graphically in Figures 27 through 29, and statistically in Tables 14.1, 14.2, and 14.3 (see Appendix H.4). Note as well that because items 101 and 181 demonstrated exceptional measurement precision (high information), the notable contribution of the remaining items was somewhat lost or misrepresented in the graphic presentations. Consequently, two additional graphs (Figures 30 and 31) with items 101 and 181 removed are also included to better illustrate that all of the remaining items contribute reliable variance to the total scale.

Inspection of the IRT results indicates that the new ANT scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable, but again high theta ranges (M = 0.10, 1.05, and 1.90 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.57, *range* = 1.17 to 2.67). The item and test information functions corroborate that the ANT items discriminate well across the mid to high trait range. The total test information function was again shifted toward the high end of theta: The percent of total test information was greater at the high trait range (66% across 0 to +3 theta range vs. 25% at 0 to -3 range). Despite several different augmentation attempts, additional

PAI items that provided either greater representation of the full trait range or strong assessment at the lower end of the ANT domain could not be identified.

As illustrated in Table 14, items that fit on the new ANT scale were drawn from several different scales on the original PAI. In particular, many items are from the original Antisocial Features, Borderline Features, Aggression, and Drug Problems scales. Evidence of content validity is strong. The items appear consistent with core APD features and appear to sufficiently reflect the full *DSM* criteria set. Compared to the original ANT scale, however, the items on this version include more representation of physically aggressive and violent behaviour. In sum, the new ANT scale demonstrates acceptable psychometric properties for screening purposes on a broadband measure. With respect to clinical utility, given that the measurement precision is strongest at the mid through high theta range, the new scale should be appropriate for clinical populations. Measurement will be less reliable for individuals who fall at the lower end of the trait continuum (< 1 *SD*).

BOR

As previously reviewed, the original PAI battery already includes a BOR features scale with four respective subscales. Consequently, exactly akin to the ANT scale modeling process, three strategies were used here in creating and analyzing the BOR scale data. First, IRT analyses were run on the original BOR total and subscale structure as defined in the manual. Second, the original items were reconfigured based on the results of the IRT analyses with a view to improve the psychometric properties of the original BOR scale. Third, an attempt was made to create an entirely new BOR scale using the same rational-empirical process employed for the other PD scales. The goal of the third approach was to explore whether the original BOR scales could be improved through augmentation with other items. Results of the three strategies follows.

Original BOR Scales. IRT analyses were run on both the original PAI BOR total scale and the respective subscale structure. Items for the original BOR total scale arranged in descending order based on the amount of information contributed to the total scale are displayed in Table 15. The original PAI subscale membership for each item is also noted.

Table 15

Original BOR Scale Items

Original PAI Subscale	BOR Item description (PAI item number)
BOR-A	(14) My mood can shift quite suddenly.
BOR-A	(54) My moods get quite intense.
BOR-I	(57) Sometimes I feel terribly empty inside.
BOR-A	(94) My mood is very steady.
BOR-I	(17) My attitude about myself changes a lot.
BOR-A	(134) I have little control over my anger.
BOR-N	(19) My relationships have been stormy.
BOR-I	(97) I worry a lot about other people leaving me.
BOR-S	(183) When I'm upset, I typically do something to hurt myself.
BOR-I	(137) I often wonder what I should do with my life.
BOR-A	(214) I've had times when I was so mad I couldn't do enough to express all my anger.
BOR-S	(223) I'm too impulsive for my own good.
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.
BOR-A	(174) I've always been a pretty happy person.
BOR-N	(99) People once close to me have let me down.
BOR-N	(139) I rarely feel very lonely.
BOR-S	(303) I'm a reckless person.
BOR-N	(179) I've made some real mistakes in the people I've picked as friends.
BOR-N	(59) I want to let certain people know how much they've hurt me.
BOR-I	(217) I don't get bored very easily.
BOR-I	(177) I can't handle separation from those close to me very well.
BOR-S	(263) I spend money too easily.
BOR-N	(219) Once someone is my friend, we stay friends.
BOR-S	(343) I'm careful about how I spend my money.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .90; total scale *H* coefficient was .34; and the individual item *H* values (H_g) ranged from .28 to .46. The alpha estimate is strong and supports that the BOR scale items demonstrate acceptable inter-item correlations. The total scalability coefficient is acceptable, but somewhat weaker than ideal and falls in the lower bound range. Inspection of the individual H_g values revealed that four of the 24 scale items fell below the acceptable range, one was moderate, and the remaining fell in the lower bound range. Although acceptable, the scalability findings are lower than expected given the alpha level and IRT results (see below). Inspection of the inter-item correlation matrix reveals that all *r*s are <.70, and the majority are moderate (.20 to .35). The individual item H_g values were less than ideal, but nonetheless acceptable (and improved from the ANT H_g values). When interpreted in context with the IRT results that follow, these findings provide modest support for the unidimensional assumption. The incongruence between the strong alpha value and lower bound total *H* value suggests that the scale may be multifaceted, but likely more cohesive than the original ANT scale. Like the ANT total scale results, this finding is also not unexpected given that the BOR scale was specifically constructed to comprise four subtests.

A reasonable fit of the GRM was obtained for almost all items. Results of the IRT analyses are displayed graphically in Figures 32 through 34, and statistically in Tables 15.1, 15.2, and 15.3 (see Appendix I.1). Inspection of the IRT results indicates that the majority of the original BOR scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -1.01, 0.45, and 1.67 at β1, β2, and β3, respectively). As well, almost all α parameters are strong (M = 1.23, *range* = 0.75 to 1.83). With respect to CCC functions, seven problematic items were

identified (p177, p263, p343, pr217, p219, p183, p303). Either the ordinal properties of these items were weak overall and/or the discrimination parameters clustered in too restrictive a range toward the high end of theta. Eight items had poor measurement precision or contributed low reliable variance. As illustrated in the test and item information functions, the remaining items show strong psychometric properties. The total test information function is smooth and reasonably well distributed across the theta levels. There is some evidence that the BOR scale provides slightly improved measurement precision across the mid to upper compared to the lower range of theta (e.g., 48% across the 0 to +3 theta range, vs. 38% across -3 to 0 range). As a clinical tool, this difference is likely not detrimental because the most pronounced drop in measurement precision occurs outside minus two standard deviations on the theta continuum. Overall, the BOR items appear to discriminate well across the full trait continuum. The results across the respective analyses indicates that the original BOR scale demonstrates acceptable psychometric properties for screening purposes on a broadband measure. Notwithstanding, some problematic items were identified, which suggests that the scale can be improved.

Results of the subscale analyses for the original BOR subscales were also positive, particularly given that each of the respective subtests contains only six items. A reasonable fit of the GRM was obtained for each subscale. The results of the IRT analyses are displayed graphically in Figures 35 and 36, and statistically in Tables 15.4 through 15.7 (see Appendix I.2). The majority of the individual item response CCCs were acceptable across all subscales. Only three items remained problematic: pr217, pr139, and pr219. Based on visual inspection of the CCCs and consideration of the statistical parameters, with the exception of one item (pr139), the psychometric properties of the items remained the same or demonstrably improved (cf., pr343, p14, p54, p303). On average, the discrimination parameters were improved for all subscales and

the response category thresholds were less extreme overall (e.g., The mean β3 threshold parameters were less extreme for all subscales). Inspection of the individual IIFs indicates that all subscales demonstrate unidimensional properties and most items contribute reliable variance to the respective subscale. Consistent with the problematic CCCs just identified, however, three items failed to contribute adequate reliable variance (pr217, pr139, pr219) to their respective subscales. Item p183 was also problematic, contributing low reliable variance overall except at the very extreme trait range (> 3.0 *SD*). Because two of the problematic items (pr139 and pr219) are both from the BOR-N subscale, the utility of this subscale is questionable because only four items appear to be contributing to the reliable variance or measurement precision of this subscale.

Overall, the test information functions were reasonably smooth and well distributed across the full trait range for each subscale. The BOR-A TIF demonstrated evidence of a mild saw-tooth pattern. Inspection of the item CCCs indicates this is attributable to two items having very strong discrimination parameters and several items showing less than equivalent measurement precision across each of the respective response categories (e.g., Item pr94: The "slightly true" response option contributes more information compared to the "mainly true" response option). And, the TIF for the BOR-S subscale indicates that the scale provides greater measurement precision across the mid to upper versus lower trait ranges (e.g., 58% at 0 to +3 vs. 31% at -3 to 0). Measurement precision is near equivalently distributed across the lower and upper trait ranges for the remaining subscales: All three subscales indicated between 40% to 50% of the total test information is distributed within both the 0 to +3 and -3 to 0 theta ranges.

Results of the additional tests of monotonicity and dimensionality for the BOR-A, -I, -N, and -S subscales were as follows: Cronbach's coefficient alphas were .80, .73, .70, and .76, and

the total scale *H* coefficients were .44, .35, .33, and .45, respectively. For the BOR-A subscale, the alpha estimate is acceptable, but low to moderate. And the total scalability estimate is moderate. Both of these coefficients, however, are quite impressive because they are derived from a scale with only six items. Consistent with the CCC results, all of the individual item H_g values were acceptable (*range* = .39 to .50). Moreover, all values were improved from the total scale analyses and fell near or within the moderate range. The convergence across all indicators suggests that the BOR-A subscale demonstrates acceptable psychometric properties: Evidence of unidimensionality and monotonicity are supported. And, given that items p14, p54, and pr94 contribute the most reliable variance, qualitative considerations of content validity suggest that the scale indeed assesses affective instability. Difficulty with affect regulation specific to anger management, another core feature of BPD, also appears to be captured by this subscale.

For the BOR-I subscale, the alpha estimate is acceptable, but low. And the total scalability estimate also falls within the lower bound range. Consistent with the CCC results, all but one (pr217) of the individual item H_g values were acceptable (*range* = .27 to .40). No marked change across these indexes compared against the full scale IRT analyses is readily apparent. Considered in context with all of the additional findings, the BOR-I subscale demonstrates minimally acceptable psychometric properties. Evidence of unidimensionality and monotonicity are supported. Because one item does not contribute reliable variance (pr217) and a second contributes only minimally (p177), the BOR-I subscale essentially consists of only four or five items. Consequently, even though the remaining items are strong from an IRT standpoint, there appears to be an insufficient number of items to generate a more reliable, overall index. Notwithstanding, as a screening tool, the BOR-I scale appears to provide adequate measurement precision across the full trait range. Further, qualitative considerations of content validity suggest

that the scale indeed assesses identity disturbance. Of interest, the results also suggest that both the "chronic emptiness" and "fear of abandonment" features of BPD are captured by this subscale. This finding is particularly important because it speaks to the potential cohesiveness of a substantial component of the BPD diagnostic criteria set.

For the BOR-N subscale, the alpha estimate is acceptable, but low. And the total scalability estimate also falls within the lower bound range. Consistent with the CCC results, four of the six individual item H_E values were acceptable (*range* = .28 to .38) and fell within the lower bound range. No marked change across these indexes compared against the full scale IRT analyses is readily apparent. Considered in context with all of the additional findings, the BOR-N subscale demonstrates minimally acceptable psychometric properties. Evidence of unidimensionality and monotonicity are supported. Because two items do not contribute reliable variance (pr139, pr219) and a third contributes only minimally (p177), the BOR-N subscale essentially consists of only three or four items. Like the BOR-I results, even though the remaining items are strong from an IRT standpoint, their appears to be an insufficient number of items to generate a more reliable, overall index. As a screening tool, the BOR-N scale appears to provide adequate measurement precision across the full trait range. And, qualitative considerations of content validity support that the scale assesses the instability of interpersonal relationships feature of BPD. Of concern, however, because potentially half of the items are contributing unreliable variance to the total subtest score, it is questionable whether the total subtest score is a valid index of functioning on this trait.

For the BOR-S subscale, the alpha estimate is acceptable, but low. The total scalability estimate falls within the moderate range. All of the individual item H_E values were acceptable and also fell within the moderate range (*range* = .43 to .49). Results across these indexes are

demonstrably improved compared against the results from the full scale IRT analyses. Considered in context with all of the additional findings, the BOR-S subscale demonstrates acceptable psychometric properties. Evidence of unidimensionality and monotonicity are supported. As previously noted, however, the BOR-S subscale provides greater measurement precision at the higher, more extreme trait range. It is unlikely that this is a significant limitation, because the measurement precision does not markedly drop until close to minus two standard deviations. Qualitative considerations of content validity suggest that the scale assesses the impulsivity feature of BPD. Indeed, impulsivity may be a more apt scale descriptor versus stimulus-seeking.

Modified Original BOR Scale. Items for the final version of the modified original BOR PD scale are displayed in Table 16, and the original PAI subscale membership for each item is also noted.

Table 16

Modified Original BOR Scale Items

Original PAI Subscale	BOR Item description (PAI item number)
BOR-A	(14) My mood can shift quite suddenly.
BOR-A	(54) My moods get quite intense.
BOR-I	(57) Sometimes I feel terribly empty inside.
BOR-A	(94) My mood is very steady.
BOR-S	(183) When I'm upset, I typically do something to hurt myself.
BOR-I	(17) My attitude about myself changes a lot.
BOR-A	(134) I have little control over my anger.
BOR-A	(214) I've had times when I was so mad I couldn't do enough to express all my anger.
BOR-N	(19) My relationships have been stormy.
BOR-I	(137) I often wonder what I should do with my life.
BOR-I	(97) I worry a lot about other people leaving me.
BOR-A	(174) I've always been a pretty happy person.
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.
BOR-S	(223) I'm too impulsive for my own good.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .88; total scale *H* coefficient was .41; and the individual item *H* values (H_g) ranged from .37 to .49. The alpha estimate is moderate, and supports that the BOR scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable and also falls in the moderate range. Inspection of the individual H_g values revealed that almost half of the items fell within the high-end of the lower bound range and the other half

fell within the moderate range. These findings support that the revised BOR scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the BOR items. Results of the IRT analyses are displayed graphically in Figures 37 through 39, and statistically in Tables 16.1, 16.2, and 16.3 (see Appendix I.3). As illustrated in Table 16, items that fit on the BOR PD scale were drawn primarily from the BOR-A, -I, and -S subscales of the original PAI. Only one item is from the BOR-N scale. Removal of low information items and reanalyses with the IRT methods appears to have improved the original scale. The remaining items are near identical to the high information items from the original IRT investigation of the total scale. Evidence of content validity is strong. The items appear to capture the core features of BPD: "...pattern of instability of interpersonal relationships, self-image, and affects, and marked impulsivity" (*DSM-IV-TR*, 2000, p. 710). Also, except for transient paranoid ideation or dissociation, the items capture the full BOR PD criteria set.

Inspection of the IRT results indicates that the BOR scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable, but somewhat high theta ranges (M = -0.64, 0.54, and 1.56 at $\beta 1$, $\beta 2$, and $\beta 3$, respectively). As well, all α parameters are strong (M = 1.50, *range* = 1.16 to 2.13). The item and test information functions corroborate that the BOR items discriminate well across the full trait continuum. And some items are clearly more informative than others (e.g., p14, p54, p57, and each contribute over 7% of the total information). There is again, however, evidence that the scale provides somewhat increased measurement precision across the mid to high compared to lower trait range as 52% of the test information falls in the upper theta range (0 to +3), versus 40% in the lower theta range (-3 to 0). Of interest, it is also noteworthy that, like the ANT

results, the briefer scale appears to provide more precise information across the full trait continuum compared to the lengthier, original BOR total scale (91% vs. 86% of the total information falls within the theta range of -3 to +3, respectively).

With respect to clinical utility, the revised BOR scale appears appropriate for the assessment of both high and low functioning individuals. Measurement will be slightly more precise across the mid through higher or more extreme trait range, which is ideal for clinical populations. Assessment of lower functioning individuals will be less reliable, but it is not a significant limitation because the measurement precision does not markedly drop until minus two standard deviations. As well, like the ANT findings, this modified BOR scale derived only from original BOR items is briefer, yet appears to provide at equivalent if not stronger assessment of the latent trait. As well, from a theoretical perspective, the IRT results are interesting. Based on the information functions, the IRT data suggest that emotional dysregulation (in particular, anger management), feelings of emptiness, poor sense of self, and self-harm behaviour are the core features of BPD. Whereas, the DSM criteria set suggests that fear of abandonment, unstable relationships, poor sense of self, and impulsivity are the primary features of BPD[2]. The caveat here, however, is that the trait estimate was achieved via self report.

New BOR scale. The last strategy undertaken within this domain was an attempt to create an entirely new BOR scale using the same rational-empirical process employed for the other PD scales with a view to explore whether the original BOR scales could be improved through augmentation with other items. Items for this last version of the BOR PD scale are displayed in Table 17, and the original PAI subscale membership for each item is also noted.

---

[2] The respective *DSM* PD diagnostic criteria sets present the respective PD features in descending rank by importance.

Table 17

New BOR Scale Items

Original PAI Subscale	BOR Item description (PAI item number)
SUI	(20) At times I wish I were dead.
DEP-C	(67) Sometimes I think I'm worthless.
SUI	(60) I've thought about ways to kill myself.
SUI	(100) I've made plans about how to kill myself.
BOR-I	(57) Sometimes I feel terribly empty inside.
BOR-S	(183) When I'm upset, I typically do something to hurt myself.
BOR-A	(54) My moods get quite intense.
BOR-A	(14) My mood can shift quite suddenly.
BOR-A	(94) My mood is very steady.
PIM	(344) I rarely get in a bad mood.
AGG-P	(61) Sometimes my temper explodes and I completely lose control.
BOR-I	(17) My attitude about myself changes a lot.
BOR-I	(97) I worry a lot about other people leaving me.
BOR-N	(19) My relationships have been stormy.
SUI	(341) Things have never been so bad that I thought about suicide.
AGG-A	(258) I have a bad temper.
AGG-A	(299) My anger never gets out of control.
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .91; total scale *H* coefficient was .42; and the individual item *H* values (H_g) ranged from .35 to .48. The alpha estimate is strong, and supports that the new BOR scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is also acceptable and falls in the moderate range. Inspection of the individual H_g values revealed that seven of the 18 items fall in the lower bound range, and the remainder fall in the moderate range (One item, p183, fell in the strong range). These findings support that the new BOR scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the new BOR items. Results of the IRT analyses are displayed graphically in Figures 40 through 42, and statistically in Tables 17.1, 17.2, and 17.3 (see Appendix I.4). Inspection of the IRT results indicates that the new BOR scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable, but again high theta ranges (*M* = -0.55, 0.56, and 1.49 at β1, β2, and β3, respectively). As well, all α parameters are strong (*M* = 1.57, *range* = 0.96 to 2.40). The item and test information functions corroborate that the BOR items discriminate well across the full trait continuum. Again, however, the TIF distribution appears moderately shifted toward the high end of theta: The percent of total test information was greater at the high trait range (58% across 0 to +3 theta range vs. 34% across 0 to -3 range). Despite several different augmentation attempts, additional PAI items that provided either greater representation of the full trait range or strong assessment at the lower end of the BOR domain could not be identified.

Visual inspection of the total TIF also suggests that the distribution is much narrower (kurtotic) than ideal. This interpretation is somewhat misleading, however, because inspection of

the individual item CCCs suggests that the narrowed distribution is due to the presence of a few items providing excellent discrimination, but within only a narrow theta range (cf., p100, p20, p60). Of note, these items assess suicidal ideation. These items function more dichotomously, are endorsed with low frequency, and capture the higher end of the trait continuum. Dropping these items indeed flattens the distribution. Based on content validity and the strength of the individual item parameters (across all indexes) for these three items, however, it appears justified to retain them on the new BOR scale. Ideally, equally strong items that captured the lower end of theta should be included to off-set the high-end bias. Unfortunately, such items could not be identified within the existing PAI item pool.

As illustrated in Table 17, items that fit on the new BOR scale were drawn from several different scales on the original PAI. In particular, many items are from the original Borderline Features, Aggression, and Suicidal Ideation scales. One item from each of the Depression and Positive Impression management scales were also represented. Evidence of content validity is strong. The items again appear consistent with core BPD features and appear to sufficiently reflect the full criteria set. Emotional lability, feelings of emptiness, anger management, and self-harm features were again prominent. Compared to the original BOR scale, these results suggest that, consistent with the *DSM* criteria set, the self-harm concept also appears to specifically encompass suicidality.

In sum, the total scale indexes for the new BOR scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. The new BOR scale is appropriate for the assessment of both high and low functioning individuals. Like the revised original BOR scale, measurement will be slightly more precise across the mid through higher or more extreme trait range, which is ideal for clinical populations. Assessment of lower

functioning individuals will be less reliable, but it is not a significant limitation because the measurement precision does not markedly drop until close to minus two standard deviations.

HIS

Items for the final version of the HIS PD scale are displayed in Table 18, and the original PAI subscale membership for each item is also noted.

Table 18

HIS Scale Items

Original PAI Subscale	**HIS Item description (PAI item number)**
SCZ-S	(270) I make friends easily.
DEP-A	(286) I'm almost always a happy and positive person.
BOR-A	(174) I've always been a pretty happy person.
WRM	(13) I'm a very sociable person.
WRM	(53) It's easy for me to make new friends.
WRM	(93) I like to meet new people.
DEP-A	(246) Lately I've been happy much of the time.
SCZ-S	(310) I keep in touch with my friends.
SCZ-S	(230) I like to be around other people if I can.
SCZ-S	(110) I'm a loner.
BOR-I	(57) Sometimes I feel terribly empty inside.
NON	(121) I spend most of my time alone.
ANX-A	(4) I am so tense in some situations that I have great difficulty getting by.
PAR-R	(77) I seem to have as much luck in life as others.
NIM	(169) People don't understand how much I suffer.
RXR	(2) I have some inner struggles that cause problems for me.
DEP-P	(115) I rarely have trouble sleeping.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .91; total scale *H* coefficient was .43; and the individual item *H* values (H_g) ranged from .34 to .50. The alpha estimate is strong, and supports that the HIS scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable and falls in the moderate range. Inspection of the individual H_g values revealed that five of 17 items fell in the lower bound range and, with the exception of item p286 which was strong (H_g = .50), the remaining items fell in the moderate range. These findings support that the HIS scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the HIS items. Results of the IRT analyses are displayed graphically in Figures 43 through 45, and statistically in Tables 18.1, 18.2, and 18.3 (see Appendix J). Inspection of the IRT results indicates that the HIS scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (*M* = -1.49, -0.37, and 0.93 at β1, β2, and β3, respectively). As well, all α parameters are strong (*M* = 1.56, *range* = 0.99 to 2.17). The item and test information functions corroborate that the HIS items discriminate well across the full range of the HIS trait levels. Note that a nonsmooth IIF distribution is noted for several of the high information items (p270, p286, p174, p13, and p53). Rather than a psychometric weakness, inspection of the individual item CCCs indicates that this pattern is attributable to the more precise discrimination parameters or sharp item response category thresholds within these items. Overall, the percent of test information is well distributed above and below the mean, however, slightly more information is available across the lower range of theta (53% within -3 to 0 and 40% within theta range of 0 to +3).

As illustrated in Table 18, items that fit on the HIS PD scale were drawn from several different scales on the original PAI. In particular, many items are from the original Schizophrenia – Social Detachment and Interpersonal (warmth) scales. Evidence of content validity with respect to assessing a single dimension is strong: The items appear to assess a domain of extraversion related to high sociability and gregariousness. Note that in modeling the HIS scale, substantial overlap with the NAR scale was encountered. To minimize the overlap, a decision was made to keep the grandiose, self important, attention demanding type of items on the NAR scale. As a consequence (and as will be evident in the next section), the HIS scale appears to capture a more normative form of extraversion or sociability/gregariousness, whereas the NAR scale may capture a more distorted, egocentric form of gregariousness.

In sum, the total scale indexes for the HIS scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. As a clinical tool, the HIS scale should adequately assess individuals in both clinical and nonclinical populations. As noted, the slight positive skew of the overall test information distribution that is discernable at the highest end of the trait distribution suggests that measurement precision will become less reliable for the assessment of individuals at the extreme high end of the trait distribution. The limitation is mild ($\geq$ 2.0 *SD*), but should be taken into consideration for diagnostic purposes.

NAR

Items for the final version of the NAR PD scale are displayed in Table 19, and the original PAI subscale membership for each item is also noted.

Table 19

NAR Scale Items

Original PAI Subscale	NAR Item description (PAI item number)
MAN-G	(268) Lately I feel so confident that I think I can accomplish anything.
DEP-C	(227) I think good things will happen to me in the future.
DEP-C	(307) I'm pretty successful at what I do.
DEP-C	(267) I have something worthwhile to contribute.
DOM	(296) People listen to my opinions.
RXR	(202) I'm comfortable with myself the way I am.
ANX-P	(193) It's easy for me to relax.
DOM	(56) I'm a natural leader.
MAN-G	(148) I have accomplished some remarkable things.
DEP-C	(67) Sometimes I think I'm worthless.
WRM	(333) I have more friends than most people I know.
MAN-G	(28) I have many brilliant ideas.
DOM	(16) I'm a "take charge" type of person.
DEP-C	(27) I feel that I've let everyone down.
MAN-G	(108) My plans will make me famous someday.
MAN-G	(188) I think I have the answers to some very important questions.
MAN-G	(68) I have some very special talents that few others have.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .88; total scale *H* coefficient was .38; and the individual item *H* values (H_g) ranged from .32 to .45. The alpha estimate is moderate to strong and supports that the NAR scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable, and falls in the low to moderate range. The individual H_g values similarly fell in the lower through to moderate range. When interpreted in context with the additional indexes

reviewed, these findings support that the NAR scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the NAR items. Results of the IRT analyses are displayed graphically in Figures 46 through 48, and statistically in Tables 19.1, 19.2, and 19.3 (see Appendix K). Inspection of the IRT results indicates that the NAR scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -1.28, 0.01, and 1.55 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.36, *range* = 0.84 to 1.98). It is noted, however, that four items (p188, p68, p108, p27) demonstrate less than ideal individual item information functions and are, therefore, less reliable. The graphic results of the CCCs suggest this is attributable to weaker individual item response category threshold parameters: The individual response category functions are less clearly differentiated and the slopes of the functions are more gradual, which suggests that measurement is less precise overall for these four items. An attempt was made to model the NAR scale without these items, however, item deletion is not a static process. Removing the weaker items caused a cascade of changes to the psychometric properties of the remaining items. As previously described, the NAR and HIS scale were developed concurrently. Concessions were made to minimize item overlap and maintain as broad representation of the full *DSM* criteria set as possible. As a result, at this stage, removal of the weakest NAR items causes undesirable changes to the psychometric properties of the remaining scale items.

Regardless, the overall test information function corroborates that the NAR items discriminate well across the full range of the NAR trait levels. The percent of test information is equally distributed above and below the mean (45% within -3 to 0, and 43% within theta range

of 0 to +3). As a clinical tool, the NAR scale should adequately assess individuals in both clinical and nonclinical populations because the scale items provide reliable information at both the high and low end of the trait dimension.

In sum, the total scale indexes for the NAR scale demonstrate acceptable psychometric properties for screening purposes on a broadband measure. As illustrated in Table 19, items that fit on the NAR PD scale were drawn from several different scales on the original PAI. In particular, many items are from the original Mania – Grandiosity and Depression – Cognitive features subscales. The additional scale items reflect themes of dominance and resistance to change. Evidence of content validity is strong: All items are clearly consistent with the *DSM-IV-TR* criteria set for NAR PD, and the full criteria set is represented. Consideration of both content validity and the quantitative findings suggests that the NAR scale could be used for applied research and clinical purposes.

DEP

Items for the final version of the DEP PD scale are displayed in Table 20, and the original PAI subscale membership for each item is also noted.

Table 20

DEP Scale Items

Original PAI Subscale	DEP Item description (PAI item number)
ANX-C	(265) I usually worry about things more than I should.
RXR	(2) I have some inner struggles that cause problems for me.
ARD-P	(66) I have exaggerated fears.
NON	(121) I spend most of my time alone.
RXR	(122) I need some help to deal with important problems.
SCZ-S	(110) I'm a loner.
SCZ-S	(150) I don't feel close to anyone.
ANX-C	(185) I don't worry about things any more than most people.
BOR-I	(97) I worry a lot about other people leaving me.
DOM	(136) I have trouble standing up for myself.
DOM	(216) I prefer to let others make decisions.
ARD-P	(106) I get very nervous when I have to do something in front of others.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .82; total scale *H* coefficient was .31; and the individual item *H* values (H_g) ranged from .26 to .35. The alpha estimate moderate, which is significant given that the scale has 12 items and supports that the DEP scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable and falls in the lower bound range. Inspection of the individual H_g values revealed that nine items were acceptable and fell within the lower bound range. The remaining three fell below the lower bound criterion. Inspection of the ICCs for the three problematic items (p106, pr185, p216) provided additional information. Because the H_g estimate is a closely related, nonparametric form of the α estimate in parametric

IRT analyses, the lower than ideal H_g estimates suggests that these scale items demonstrate weaker discrimination across many of the respective response category thresholds. Review of the α estimates and visual inspection of the graphic results for these items corroborates this interpretation (see IRT section below). Although the discrimination power of each response category threshold for these items was less than ideal, an ordinal property across the category thresholds was readily discernable, and all difficulty and discrimination parameter estimates fell within acceptable limits for these items. Moreover, additional items with an improved fit (re IRT and nonparametric indexes) could not be identified from the existing PAI item pool. Overall, considered in context with all of the available psychometric indexes, the scalability findings support that the DEP scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the DEP items. Results of the IRT analyses are displayed graphically in Figures 49 through 51, and statistically in Tables 20.1, 20.2, and 20.3 (see Appendix L). Inspection of the IRT results indicates that the DEP scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -1.01, 0.31, and 1.61 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.24, *range* = .90 to 1.52). The item and test information functions corroborate that the items assess a broad range of the DEP trait levels. Measurement precision is strongest through the midrange. The TIF is near symmetrical and shifted slightly to the right. This indicates that somewhat more information is available toward the high compared to low end of the DEP continuum (48% within theta range of 0 to +3, and 39% within -3 to 0). Measurement precision remains reliable for the assessment of individuals in the highest trait range (28% > 1 *SD*) and only slightly less reliable at the lowest trait range (21% < 1 *SD*).

In sum, the total scale indexes for the DEP scale results demonstrate acceptable psychometric properties for screening purposes on a broadband measure. As illustrated in Table 20, the items that fit on the DEP PD scale were drawn from a variety of subscales of the original PAI. Evidence of content validity is strong: All items are consistent with the *DSM-IV-TR* criteria set for DEP PD. Consideration of the item content suggests that the DEP scale comprehensively captures key DEP PD features, including exaggerated fears/separation anxiety, deference or submissiveness, and lack of assertiveness. Less well captured are more specific behavioural consequences of the negative affective and cognitive features (e.g., "excessive lengths to obtain nurturance and support from others", *DSM-IV-TR*, 2000, p. 725). Overall, consideration of both the content validity and the quantitative findings suggests that the DEP scale can be used for applied research and clinical purposes.

COM

Items for the final version of the COM PD scale are displayed in Table 21, and the original PAI subscale membership for each item is also noted.

Table 21

COM Scale Items

Original PAI Subscale	COM Item description (PAI item number)
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.
ANT-S	(119) My behaviour is pretty wild at times.
ANT-A	(131) I used to lie a lot to get out of tight situations.
BOR-S	(263) I spend money too easily.
PIM	(264) I sometimes make promises I can't keep.
AGG-P	(61) Sometimes my temper explodes and I completely lose control.
STR	(322) My life is very unpredictable.
BOR-N	(19) My relationships have been stormy.
ANT-A	(91) I've done some things that weren't exactly legal.
BOR-A	(54) My moods get quite intense.
BOR-S	(343) I'm careful about how I spend my money.
ARD-O	(45) I have impulses that I fight to keep under control.
WRM	(332) I'm very impatient with people.
ANT-A	(291) I've never taken money or property that wasn't mine.
ANT-S	(239) I like to drive fast.
ANT-A	(251) I've never been in trouble with the law.

Note. Items are arranged in descending order based on amount of information contributed to the total scale.

Results of the initial tests of monotonicity and dimensionality were as follows: Cronbach's coefficient alpha was .86; total scale *H* coefficient was .34; and the individual item *H* values (H_g) ranged from .25 to .41. The alpha estimate is moderate and supports that the COM scale items demonstrate acceptable inter-item correlations. The total scale *H* coefficient is acceptable, and falls in the lower bound range. Inspection of the individual H_g values revealed that with the exception of two items (p251 and pr239), the scalability values fell within the mid

to lower, but acceptable range. These findings support that the COM scale demonstrates acceptable monotonicity and unidimensional properties.

A reasonable fit of the GRM was obtained for the COM items. Results of the IRT analyses are displayed graphically in Figures 52 through 54, and statistically in Tables 21.1, 21.2, and 21.3 (see Appendix M). Inspection of the IRT results indicates that the COM scale items have acceptable item response CCCs: Within each item, the category thresholds demonstrate an ordinal property and fall within acceptable theta ranges (M = -1.53, -0.43, and 0.74 at β1, β2, and β3, respectively). As well, all α parameters are strong (M = 1.28, *range* = 0.79 to 1.98). The item and test information functions indicate that the COM items discriminate well across the mid to lower range of the COM trait levels. The test information function distribution is shifted slightly to the left, which suggests that the scale is more discriminating at the lower trait ranges. Overall, 52% of the test information falls within -3 to 0, and 37% within the theta range of 0 to +3. As a clinical tool, the COM scale should adequately assess individuals in nonclinical populations and mild to moderate clinical ranges (≤ 2 SD above the mean). Although a more balanced representation of items across the full range is desirable, additional PAI items that assessed the same COM dimension, but at the highest trait levels could not be identified from the existing pool.

As illustrated in Table 21, most items that fit on the COM scale are reverse coded items from the original PAI Antisocial and Borderline scales. Of interest, this scale was difficult to model with IRT. Because many items with high prototypicality ratings that well captured the full *DSM* criteria set for COM PD were available in the original PAI item pool, it was anticipated that a comprehensive, psychometrically sound COM scale would be created. Unexpectedly, the initial fit of the IRT model to the preliminary COM scale was poor. In the initial modeling

process, there was no evidence of unidimensionality, item signs or directionality (-/+) conflicted in unpredictable ways, and the amount of information from any single or small cluster of items was poor. Moreover, items that captured hallmark COM features (e.g., perfectionism, attention to detail) repeatedly failed to fit[3]. Following the iterative construction process, a group of items that adequately fit the GRM was identified. The resulting COM scale appears to reflect a dimension of inhibited or reserved psychosocial and behavioural functioning. Evidence for content validity is moderate: The items appear to capture the passive quality of COM (e.g., mental and interpersonal control), but less adequately capture the more active features of COM (e.g., exceedingly high standards, productivity, and conscientiousness).

Additional Convergent and Discriminant Validity Evidence

Interscale correlations among the new PAI PD scales and with the MCMI-III PD scales were run (see Appendix N). Consistent with extant PD research, the correlations were relatively moderate to strong overall (e.g., Morey et al., 1985). This is not unexpected given the noted conceptual overlap. Of greater importance is the pattern (direction and relative strength) of the correlations (Widiger & Coker, 1992). The pattern of interscale correlations for the PAR, SZD, and SZT scales largely fell in the expected directions: These scales correlated more strongly with each other and the internalizing related domains (AVD, DEP) and more weakly with the externalizing related domains (HIS and NAR). The scales also correlated negatively with the COM scale, which is presumably due to COM pulling high conscientiousness. PAR, SZD, and SZT also correlated strongly with the BOR scale, which is likely due to shared neurotic features related to anger or hostility, as well as shared difficulty with interpersonal relationships.

[3] As an aside, out of curiosity, IRT analyses were run on several different clusters of PAI items that appeared to more directly reflect *DSM* criteria for COM PD. No conceivable configuration of such items generated acceptable fit indexes.

The three versions of the BOR scales correlated strongly with each other, and demonstrated the same pattern of correlations with the other new PD scales. The three versions of the ANT scale also correlated strongly with each other, and demonstrated the same pattern of correlations with the other new PD scales. The new BOR and new ANT scales correlated more strongly with each other compared to the original versions of these scales. The new ANT scale demonstrated the strongest (and negative) correlation with the COM scale, which is consistent with the ANT features of low conscientiousness or poor respect for rules and regulations, and high excitement seeking. The BOR scale also demonstrated the strongest, negative correlation with the COM scale. This is consistent with the low anxiety and low neuroticism features of COM compared to BOR. The BOR scales were negatively correlated with the HIS and NAR scales, which is consistent with the positive mood, optimism, low anxiety, and stronger interpersonal relationship features of NAR and HIS compared to BOR.

Overall, the pattern of interscale correlations between the new PD scales and the corresponding PD scale on the MCMI-III predominantly fell in the expected directions. Each new scale correlated moderate to strongly with its respective counterpart on the MCMI-III, and was typically the strongest correlation demonstrated. If the matched PD scale was not the strongest, a cluster of strong correlations was evident among PDs with shared features. For example, strong correlations were demonstrated across the PAI and MCMI-III PD scales that share features of high anxiety, avoidance, and fearful behaviours (e.g., AVD, DEP, BOR), which also demonstrated weak correlations with the MCMI-III PD scales that capture theoretically unrelated domains including, high conscientiousness and self-confidence (e.g., COM, NAR). In sum, although the intercorrelation pattern was not perfect, the direction and magnitude is consistent with extant PD research.

Discussion

The primary purpose of this test development exercise was to enhance the clinical utility of the PAI, an existing broadband measure of psychopathology. A rational-empirical strategy was employed to attempt to create 10 new scales that specifically assess the 10 PDs as defined in the *DSM-IV-TR*. Particular emphasis was placed on the application of IRT methodology to facilitate this process. The IRT methods employed in this project generated substantial amounts of data. As a result, several scale specific and also more general insights and conclusions were discerned. Given the measurement hurdles of conceptually overlapping, polythetic diagnostic criteria sets; less than ideal theoretical foundation; and the constraint of unidimensional statistical modeling, the results of the overall scaling process were very encouraging. It is clear from the IRT findings that the PAI items can be successfully reconfigured into additional, narrow scales. Convergence across most of the parametric IRT and nonparametric coefficients provided satisfactory to strong structural validity evidence.

The combined rational-empirical approach yielded psychometrically sound, new narrow scales for each of the 10 *DSM* PDs. Results of stage one, the rational component, indicated that the PAI items could be rearranged to capture all 10 *DSM* PDs. Thus, sufficient breadth of content was available in the existing item pool to proceed with the scale construction process. Raters demonstrated strong agreement in identifying items that capture a given PD (all mean intraclass *r*s were $\geq .80$). The scales were then subjected to IRT based analyses. Items were ultimately retained or excluded on the basis of goodness of fit with Samejima's GRM. This method enabled both individual item and total scale psychometric properties to be evaluated. For triangulation purposes, CTT and nonparametric analyses were also run. Results of the CTT based internal consistency reliability estimates (coefficient alphas) were acceptable for all new scales (αs

ranged from .82 to .94), and the magnitude was comparable if not superior to the alpha estimates reported by Morey (1991) for the original scales (αs ranged from .66 to .94). Preliminary, CTT based convergent and discriminant validity evidence was also supportive. The pattern (directionality and magnitude) of interscale correlations between the new PD scales and the MCMI-III PD scales was consistent with extant PD research. The nonparametric results or scalability coefficients were also acceptable for all total scales (all H coefficients were $\geq .30$). With the exception of three items from the COM scale, all individual item scalability coefficients were also satisfactory (most H_g values were $\geq .30$). These combined results support that each of the 10 new scales assess a single, underlying trait. Moreover, these results support that the new scales capture a range of functioning that varies by severity or intensity on each latent trait or PD.

Together, the CTT based indexes and the nonparametric and scalability results demonstrated strong convergence. Although useful, however, in the broadest sense of scale development these respective indexes provided somewhat limited psychometric information. As intended, however, the IRT results enabled much more precise investigation into the nature of respondent behaviour. The IRT results revealed that all items demonstrate ordinal measurement properties across each response option. This indicates that the graded response options function as intended across most items: It is reasonable to conclude that respondents who endorse, for example, the "slightly true" option on a given item possess less severe target trait symptoms compared to respondents who endorse the "very true" option for the same item. Also identified through this process, some items were noted to have more dichotomous versus polytomous measurement properties (e.g., Antisocial behaviour items or ANT-A items). Items that did not meet these standards were eliminated.

The IRT analyses also permitted the identification of measurement precision as a function of severity of the underlying trait for each total scale. This is compared to the CTT derived alpha indexes, which provide a composite estimate of the overall internal consistency reliability of a total scale regardless of trait severity. Consideration of the IRT derived item information functions enabled items that failed to provide satisfactory discrimination across respondents as a function of trait severity to be identified and eliminated. Consequently, each item on the new scales can be assumed to directly contribute to the ability of the total scale to accurately discriminate higher from lower trait individuals on the target PD. Further, the item information functions illustrate the specific trait range where each item provides maximal sensitivity or maximal discrimination between high and low trait respondents.

It is desirable to have highly discriminating items across as broad a range of the target trait as possible. The IRT derived test information functions illustrate that all of the new scales demonstrate acceptable measurement across a wide range of trait functioning. It was noted that each new scale demonstrated the strongest measurement precision across the mid range of their respective trait continua. Of relevance for clinical applications, all scales except COM provided acceptable measurement precision through the higher trait ranges as well. Fewer scales demonstrated acceptable measurement precision at the lowest trait ranges, which is less problematic for clinical applications. Moreover, when measurement precision dropped, it typically occurred outside one to two standard deviations above or below the mean.

These findings are not unexpected. Enhanced measurement precision across the midrange is the result of most psychological phenomena being normally distributed: More data points are available to estimate the respective statistical parameters in the midrange of the trait continua. In turn, measurement accuracy is improved (Lord, 1980). Consequently, although each subscale

demonstrates enhanced measurement precision across the midrange, it is not necessarily a weakness of the scale per se that contributes to the loss of precision at the extreme ends of theta (although the measurement limitation remains). Given that enhanced clinical utility is the primary goal of this scale development exercise, the capability of the new scales to adequately assess functioning within the mid through to higher or clinical trait ranges is promising for future research and clinical applications.

The specific IRT results for two narrow scales are discussed here in greater detail. Results generated from the ANT and BOR scales require more in depth discussion because a narrow scale (and subscales) for each already exists on the original PAI. Consequently, additional considerations are needed to determine which version of the scales to recommend for future applications. ANT results will be reviewed first, followed by BOR. It is difficult to determine whether or not to recommend adoption of the new ANT scale that was derived through the same rational-empirical approach employed for the other PDs or to recommend continued use of the original APD scale and respective subscales. Because the scales for the APD domain have a developing literature that largely supports their psychometric properties and the forensic community has widely adopted the existing measure, a convincing argument based on empirical evidence is needed to justify adoption of a new item configuration.

Results of the original ANT composite scale indexes were minimally acceptable (lower bound range), which appears attributable to both problematic individual items and lack of unidimensionality. Given that the original scale was intended to be multidimensional, it was noted that weak evidence for unidimensionality was not necessarily problematic. Notwithstanding, stronger evidence of conceptual cohesiveness across all of the items would have been preferred because the subscales are intended to capture facet components of a larger,

unified domain. Further, over half of the ANT items as configured on the original, full scale failed to demonstrate minimal acceptable standards for monotonicity and unidimensionality as measured by the nonparametric scalability indexes (H_g values). In addition, the IRT results revealed that between six and nine of the original 24 items contribute the majority of the test information on the original ANT scale. All of these findings are problematic because it suggests that individuals who score high on this scale likely comprise an overly heterogeneous group.

Morey (1991) describes that elevations on the original ANT total scale are indicative of increasing levels of antisocial behavior from mild impulsivity and risk taking behavior through moderate egocentricity and lack of empathy for others, to more pervasive exploitation of others and ultimately, diagnostic features of ANT PD. The IRT results, however, suggest it is unlikely that a single personality trait or psychopathology variable underlies responding on this scale. Further, counter to Morey's recommended scale score interpretation guidelines, the IRT results suggest that total scale scores do not follow a unidimensional, monotonic increasing function. Hence, low, medium, and high scorers on the ANT total scale cannot be assumed to demonstrate progressively increasing levels of psychopathology on a single, ANT dimension.

Consideration of the ANT subscale data provided further insights. The IRT and nonparametric indexes all indicate that the ANT-A subscale reliably assesses a single dimension across the mid to high range of the trait spectrum. Indeed, the results were remarkably strong for an eight item scale. Consistent with Morey's original intent (1991), the scale appears to capture engagement in rule and law-breaking activity. IRT results for the ANT-E and ANT-S subscales proved more problematic to interpret. The IRT results identified a few items on each scale that had strong psychometric properties, but an insufficient number to yield acceptable total scale reliability indexes. Because each subscale is brief (eight items), inclusion of any problematic

items has marked, detrimental effects on the composite statistical indexes. Conversely, the presence of three exceptionally strong items on the ANT-S scale appeared to have somewhat artificially inflated the total scale quantitative indexes. Each scale nonetheless demonstrated evidence of mild convergence toward a single dimension. The ANT-E scale appears to measure a circumscribed aspect of egocentricity or self-serving, opportunistic behaviour in individuals in the mid-upper through extremely high trait range. The ANT-S scale appears to reliably measure some form of careless, thrill-seeking behavior in the mid through high trait range. As is, however, the IRT results do not support interpretation of these subscales.

Of interest, the pattern of IRT results across the ANT subscale data is nonetheless consistent with the extant literature on self-report personality test data for the antisocial/ psychopathy spectrum construct. As reiterated in recent findings, it is routinely documented that self-report data for the assessment of the interpersonal and affective qualities of antisocial/psychopathic spectrum dysfunction (e.g., emotionality, empathy, beliefs, values, motivations) are notoriously problematic (low reliability and low validity). Whereas, self-report data for the assessment of the more objective indexes of antisocial related behaviours (e.g., stealing, forgery) typically demonstrate higher reliability and validity estimates (e.g., Edens et al., 2000).

Because the IRT based investigation of the psychometric properties of the original ANT total and subscale item configuration suggested there was room for improvement, consideration of two new item configurations for the ANT domain were warranted. The first strategy involved modifying the original ANT scale by eliminating poor functioning items on the basis of the IRT results. This approach yielded improved psychometric indexes and a greater conceptual understanding of the underlying trait. The CTT derived index remained essentially unchanged,

but the nonparametric, scalability indexes demonstrably improved. Review of the item content indicated that, with the exception of physical aggression, the brief format or modified original ANT scale reliably captures diagnostic features of the full *DSM* criteria set for ANT PD.

The second strategy involved creating an entirely new ANT scale through applying the rational-empirical approach used to construct the other PD scales. The resulting psychometric indexes were acceptable, but not demonstrably improved from the modified original ANT results. The most salient difference between these strategies was that the modified original ANT scale primarily captures thrill-seeking behavior, and the new scale appears to more strongly capture physical aggression and violence. Given that the psychometric properties are essentially equivalent and that the full *DSM* criteria set is well represented on both scales, it is interesting that the two approaches yielded different item content.

Considered together, it is difficult to recommend one ANT scale over another. Each has strengths and weaknesses. One interpretation that does appear clear and consistent across the respective findings is that the ANT construct, as defined in the *DSM*, is likely multifaceted. Perhaps even more importantly, the results suggest several areas of exploration for future theory testing. In particular, the results suggest that the domains of law-breaking, impulsivity, thrill-seeking, aggression, lack of empathy, drug and alcohol abuse, etcetera, may be very strong, unidimensional traits. The results here do not convincingly demonstrate that it is appropriate to cluster these traits under one superordinate, ANT rubric. Alternatively, if classified under the rubric of antisocial functioning, it does not appear appropriate to infer that a single underlying personality trait explains the respective behaviours or psychopathology.

Similarly, the *DSM* ANT rubric may be too heterogeneous of a clinical category to be useful for psychiatric diagnostic purposes. Results across the composite ANT scale attempts

demonstrate that largely equivalent scales can be created within this domain. Taking into consideration that the original PAI has additional clinical scales, it may prove more useful to use the modified original versus the newly created scale. Because aggression, drug, and alcohol use subscales already exist on the PAI[1], and because the original ANT scale has a developing literature with seemingly strong endorsement from the forensic community (e.g., Piotrowski, 2000), the new scale may prove somewhat redundant and likely meet with resistance. Most importantly, use of the modified original scale should, at minimum, ensure that a unidimensional ANT domain is captured. This will enhance score interpretation.

In comparison, the IRT results for the original BOR scale were superior overall. Because the original PAI also includes BOR PD scales, a decision had to be made regarding whether or not to recommend adoption of a new BOR scale/s or continued use of the existing scales. As with the ANT scales, a convincing argument based on empirical evidence is seemingly needed to justify adoption of a new item configuration. IRT results for the BOR scale and individual item analyses were generally supportive of the original total and subscale item configurations. Specifically, results for the original BOR composite scale indicate that it assesses the full trait continuum with acceptable measurement precision. The full *DSM* criteria set appears well captured, and all items contribute reliable variance to the total scale. The contribution of a small minority of items, however, was minimal and coincided with problematic intra-item psychometric properties (e.g., low discrimination, poor ordinal measurement properties). Thus, the original BOR scale demonstrates acceptable psychometric properties and, for screening purposes, appears to adequately capture the full *DSM*, BOR PD diagnosis. Results suggest that scores for both high and low scoring individuals can be interpreted. Notwithstanding, the original

[1] Note that IRT analyses were also run on these clinical subscales. The results support the structural validity of the respective scales, in particular, unidimensional trait functioning.

scale also appears to contain several items that may prove extraneous, as their inclusion does not appear to appreciably contribute to improved measurement precision.

The impact on the reliability of the total BOR scale after removing the poorer functioning items was assessed. Elimination of the psychometrically weaker items yielded a comparably strong scale from both a quantitative and qualitative perspective. The diagnostic breadth of content remained comprehensive, and all of the quantitative psychometric indexes were equivalent, if not improved. Hence, it appears that the original BOR composite scale can be improved by eliminating weaker items without sacrificing breadth of construct representation. Like the ANT results, the IRT analyses again demonstrate that a briefer scale can provide equivalent, if not stronger assessment of a given construct.

Noteworthy as well, is the primacy of mood symptoms or the affective instability feature of BPD on both the brief and original BOR scale. The DSM identifies fear of abandonment, interpersonal chaos, and identity disturbance as the three primary features of BPD. Indeed, affective instability ranks sixth. With respect to BOR theory, however, this finding is consistent with an extant literature that proposes BPD may share many features with bipolar spectrum mood disorders (e.g., Stone, 2006) or, at minimum, speaks to diagnostic overlap. This is a controversial area of research (cf., Paris, 2007; Paris, Gunderson, & Weinberg, 2007; Pies & MacKinnon, 2007), and this finding highlights the marked utility of IRT methods in the domain of psychopathology assessment and theory testing.

In addition, Morey (1991) does not advocate interpretation of low scores on the original BOR total scale and describes midrange scores as indicating emotional and interpersonal stability. Elevations are reportedly indicative of increasing levels of personality psychopathology. Specifically, elevations suggest progressive decompensation from minor

moodiness, sensitivity, and general uncertainty; into increasing anger and interpersonal difficulties; followed by more pronounced emotional lability, impulsivity, neediness, and an inability to maintain relationships; and ultimately, diagnostic features of BPD. Compared to ANT, the IRT results for the BOR scale are more consistent with a unidimensional, monotonic increasing function. In turn, as Morey intended, medium and high scorers on the BOR total scale can be assumed to demonstrate progressively increasing levels of psychopathology on a single, BOR trait. Moreover, the results indicate that Morey's dimensional conceptualization can be extended to incorporate the lower range as well. Low scores can be assumed to indicate low levels of the same BOR trait.

The subscale item configuration of the original BOR scale was also assessed. On average, the results support the original subscale structure. Indeed, a particularly notable finding was that many of the IRT derived individual item parameters were improved when the items were analyzed in the subscale, versus total scale configuration. This supports the facet domain conceptualization of the composite BOR scale. Further, the respective BOR subscale results were remarkably strong for measurement scales with only six items. Measurement precision for the BOR-A subscale was the strongest. It assesses the affective instability feature of BOR, in particular, general lability and anger management. The cohesiveness of this scale is noteworthy with respect to theory. It supports that the mood dysregulation component of BOR extends to the realm of anger regulation difficulties, as opposed to only encompassing a more two dimensional, euphoric versus dysthymic spectrum difficulty.

The remaining three subscales, BOR-I, -N, and -S, capture the identity disturbance, feelings of emptiness, and fear of abandonment; interpersonal relationship difficulties; and impulsivity and self harm components of BOR PD, respectively. The IRT results for these scales

were acceptable, but low-moderate compared to the BOR-A results. Each subscale contains problematic items, which poses significant measurement difficulties on scales with only six items. Nonetheless, the IRT and nonparametric results indicate that an acceptable clustering of the majority of items on a single, dominant trait was clearly discernable for each subscale. Moreover, measurement precision held for a substantial range of the respective trait continua, both above and below the mean for each of the subscales. Overall, although there is room for improvement, the subscale structure of the original BOR scale demonstrates acceptable psychometric properties. The IRT results support applied use of the subscales across both clinical and nonclinical populations, and suggests that both high and low scores can be interpreted. Lastly, the IRT results will be particularly useful for informing the next edition of the PAI battery. Because the subscales definitely appear to capture a single domain that conceptually fits with BOR PD theory, but problematic items were nonetheless identified, these few items can be specifically targeted for revision (e.g., rewording or replacing).

Although both the composite and subscale structure of the original PAI was largely supported, as just discussed, the IRT method identified areas where measurement could be improved. The last round of analyses, therefore, explored whether measurement could be enhanced by augmenting the existing composite scale with other items from the original PAI item pool. Results of this series of IRT analyses yielded comparable psychometric evidence: The parametric and nonparametric indexes were not substantially improved. As well, the new scale pulled many of the original BOR items, which further speaks to the validity of the underlying trait. Of interest, however, the new scale also pulled several non-BOR items. Evaluated from a purely quantitative framework, the new scale is not demonstrably improved from the modified original BOR. Seemingly more important, however, is that the new item configuration identified

several noteworthy findings with respect to BPD theory and directions for future research. Consistent with the literature on BPD, suicidal ideation and self harm were primary on the new configuration. Given that the intent of the original PAI is a broadband measure of personality and psychopathology, Morey (1991) had the forethought to include suicidality items. It is significant (and likely not surprising to practicing clinicians) that several suicidal ideation items specifically clustered with the other hallmark BOR features (low mood, emotional lability, and insecure identity) on a unidimensional trait continuum. Moreover, as a manipulation check, IRT analyses were run on just the suicide subscale items, and evidence in support of a strong, unidimensional domain was demonstrated. Thus, the convergence here of several suicidality items with the BOR items, lends theoretical support to the BOR PD trait construct as delineated in the *DSM*.

With respect to practical utility, it remains difficult to recommend one of the BOR scales over another because evidence across the quantitative and qualitative indicators is comparable. Given that the existing item configuration is consistent with *DSM* theory; the subscale structure demonstrated acceptable psychometric properties; interpretative norms have already been established, and an empirically based literature continues to amass, it appears justified to continue to use the original BOR scales. The IRT analyses, nonetheless, demonstrate that equally viable and briefer alternatives exist, and also highlight areas for future study.

Considered together, the review of the BOR and ANT findings highlight many of the additional conclusions and insights that were discerned through this scale construction process. First, using content validity as one proxy to estimate construct validity, results of the current investigation revealed additional strengths and some limitations across the newly derived scales. From a very stringent perspective of content validity, some of the newly constructed scales did not comprehensively address the full diagnostic criteria set for each respective *DSM* PD. This

limitation requires careful consideration. Although comprehensive assessment is ideal, it is not a feasible goal for a broadband measure. As noted by Widiger and Coker (2002), it would take hours to comprehensively assess all ten PDs in a single sitting. The objective of this investigation was to create PD subscales that meet, and preferably exceed, minimum psychometric standards for screening purposes on a broadband measure. Further, the objective was to increase the clinical utility of an existing measure, which unavoidably constrains the universe of available items. There is also substantial conceptual overlap across many of the diagnostic criteria sets (Costa & Widiger, 2002; Jablensky, 2002). Considering these factors, the expectation that the full *DSM* criteria set be comprehensively represented on each scale is unrealistic. Notwithstanding, despite these challenges, each new PD scale must nonetheless demonstrate sufficient breadth of content coverage and measurement specificity that the captured domain can reasonably be interpreted as the target PD. Direct assessment of this premise is beyond the scope of this study. This assumption will need to be more explicitly tested in follow up external validity investigations.

Second, consistent with the sentiment of longstanding advocates for more widespread application of IRT methods, the results overwhelmingly demonstrate the utility of applying IRT methods in the domain of psychopathology assessment - for both evaluating existing and constructing new measures. As previously discussed, several authors who work with IRT (e.g., Embretson & Reise, 2000) have routinely emphasized that psychologists have been very slow to adopt IRT based methods in measurement endeavors (outside psychoeducational testing). This study demonstrates that reliance on CTT methods contributes to a loss of measurement information that is otherwise available. When the measurement values of CTT derived indexes (e.g., correlations, coefficient alpha) on this project were compared against the IRT indexes (e.g.,

CCCs, IIFs, TIFs), the amount of additional information derived from the IRT methods was substantially improved.

Third, through cross comparing the movement of numbers on the various CTT and IRT indexes during the construction process and from the perspective of someone new to IRT, I became overwhelmingly convinced of the theoretical and clinical utility of IRT methods in the psychopathology domain. The theta estimate indeed appears to capture some form of unitary trait functioning that cannot be explained away by item endorsement frequency, response category frequency, or simple inter-item correlation patterns. Unfortunately, the IRT methods cannot identify with any more certainty than traditional CTT methods what exactly the underlying trait is. Nonetheless, the ability to say with confidence that a core variable has been identified that very precisely explains a particular endorsement pattern across several items on a measure of psychopathology or personality suggests that measurement of psychopathology constructs can be markedly improved with more widespread application of this technique.

This run of analyses alone revealed several useful applications. For example, akin to the MMPI/2 related profile interpretation approach, Morey has stated that he is interested in expanding the interpretive applications of the PAI through creating additional "configural profiles" analyses (1991, p. 21). He advocates use of factor analytic based methods to identify or create the various profiles. Results of this study suggest that IRT methods may prove more useful for this purpose. The results here demonstrate that the PAI items can indeed be reconfigured in alternative ways that seemingly capture psychometrically sound and conceptually meaningful scales or profiles. Anecdotally, several psychometrically sound and conceptually interesting scales (from both a clinical and theoretical perspective) were also identified during the iterative scale construction process. This observation suggests that, in addition to PD diagnoses, IRT

methods should prove useful in identifying alternative composite indexes or item profiles on the PAI. Furthermore, the results demonstrate that IRT based methods can provide more precise information about the contribution of each item to the measurement of a respective domain and also the range of functioning within that domain that the entire scale is able to assess. This information is not available with factor analytic techniques.

Lastly, an additional insight revealed through the IRT process that can be followed up in future studies is length of testing. On several occasions, the construction process demonstrated that fewer items are needed to assess a single domain of functioning than CTT methodology might suggest. For example, the BOR and ANT results demonstrate that less than ten items that clearly tap the same functional domain can generate surprisingly sound measurement scales. As previously discussed, broadband measures are highly favored by clinicians (Piotrowski, 1999), but respondents can be overwhelmed by such long tests. Testing brevity seems to be a necessary, practical consideration. The IRT results here (BOR, ANT) have already demonstrated that (a) psychometrically weak items exist on the original PAI scales, and (b) elimination of the weaker items can generate equivalent, if not stronger scales. Hence, the results suggest that through IRT strategies, it may be possible to abbreviate the PAI without losing breadth of content representation or measurement precision.

Future Research

Results of the current study are only the first phase of a scale construction exercise. External validity testing investigations will need to follow. Suggestions for future research include: Additional investigation of the invariance of item parameters: Do the IRT results demonstrated here hold across other samples? Differential item functioning investigations can be

run to explore any evidence of item or scale bias in various subpopulations (e.g., divergent group analyses including, gender, culture, language, socioeconomic status, etc.). Lastly, predictive validity investigations are needed to investigate the utility of the new scales in aiding clinical diagnoses (e.g., signal detection designs). Demonstrating sufficient external validity evidence is a comprehensive process that requires comparison across several independent investigations over time. The emphasis of this scale construction procedure on the structural component of construct validity evidence gathering is a necessary precursor.

References

Akiskal, H. S. (1994). The temperamental borders of affective disorders. *Acta Psychiatrica Scandinavica, 89*, 32-37.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.

Allport, G. W., & Odbert, H. S. (1936). Traitnames. A psycho-lexical study. *Psychological Monographs, 47 (211)*, 171.

Alterman, A., Zaballero, A., Lin, M., Siddiqui, N., Brown, L., Rutherford, M., & McDermott, P. (1995). Personality Assessment Inventory (PAI) scores of lower-socioeconomic methadone maintenance patients. *Assessment, 2*, 91-100.

American Psychiatric Association. (1980). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd ed. revised). Washington, DC: Author.

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (Text revision). Washington, DC: Author.

Baker, J. G., Rounds, J. B., & Zevon, M. A. (2000). A comparison of graded response and Rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistics, 25(3)*, 253-270.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Bell-Pringle, V. J., Pate, J. L., & Brown, R. C. (1997). Assessment of borderline personality disorder using the MMPI-2 and the Personality Assessment Inventory. *Assessment, 4(2)*, 131-139.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57(6)*, 717-726.

Benjamin, L. S. (1993). Dimension, categorical, or hybrid analyses of personality: A response to Widiger's proposal. *Psychological Inquiry, 4(2)*, 91-95.

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4(1)*, 14-19.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, 17(1), 10-17.

Blackburn, R., Donnelly, J. P., Logan, C., & Renwick, S. J. (2004). Convergent and discriminative validity of interview and questionnaire measures of personality disorder in mentally disordered offenders: A multitrait-multimethod analysis using confirmatory factor analysis. *Journal of Personality Disorders, 18(2)*, 129-150.

Blais, M. A., Benedict, K. B., & Norman, D. K. (1998). Establishing the psychometric properties of the DSM-II-R personality disorders: Implications for DSM-V. *Journal of Clinical Psychology, 54(6)*, 795-802.

Blais, M. A., & Norman, D. K. (1997). A psychometric evaluation of the DSM-IV personality disorder criteria. *Journal of Personality Disorders, 11(2)*, 168-176.

Blashfield, R. K. (1993). Variants of categorical and dimensional models. *Psychological Inquiry, 4(2)*, 95-98.

Boccaccini, M. T., & Brodsky, S. L. (1999). Diagnostic test usage by forensic psychologists in emotional injury cases. *Professional Psychology: Research and Practice, 30(3)*, 253-259.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bolt, D. M., & Lall, V. F. (2003). estimation Of Compensatory and noncompensatory multidimensional item response models using Markov chain monte carlo. *Applied Psychological Measurement, 27*, 395-414.

Boone, D. (1998). Internal consistency reliability of the personality assessment inventory with psychiatric inpatients. *Journal of Clinical Psychology, 54(6)*, 839-843.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11(5)*, 515-524.

Bornstein, R. F. (1998). Reconceptualizing personality disorder diagnosis in the DSM-V: The discriminant validity challenge. *Clinical Psychology: Science and Practice, 5*, 333-343.

Boyle, G. J. (1996). Psychometric limitations of the Personality Assessment Inventory: A reply to Morey's (1995) rejoinder. *Journal of Psychopathology and Behavioral Assessment, 18(2)*, 197-204.

Boyle, G. J., & Lennon, T. J. (1994). Examination of the reliability and validity of the Personality Assessment Inventory. *Journal of Psychopathology and Behavioral Assessment, 16(3)*, 173-187.

Boyle, G. J., Ward, J., & Lennon, T. J. (1994). Personality Assessment Inventory: A confirmatory factor analysis. *Perceptual and Motor Skills, 79*, 1441-1442.

Brannic, M. (2001). http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration, scoring, and interpretation* (Rev. ed.). Minneapolis, MN: University of Minnesota.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47*, 87-111.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31(2)*, 141-154.

Chapman, L. J., Chapman, J. P., Kwapil, T. R., Eckblad, M., & Zinser, M. C. (1994). Putatively psychosis-prone subjects 10 years later. *Journal of Abnormal Psychology, 103(2)*, 171-183.

Cattell, R. B., & Cattell, H. E. (1995). Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement, 55(6)*, 926-937.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M., & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 depression scale. *Assessment*, 7, 37-54.

Clark, L. A. (1993a). *Manual for the Schedule for Nonadaptive and Adaptive Personality.* Minneapolis, MN: University of Minnesota.

Clark, L. A. (1993b). Personality disorder diagnosis: Limitations of the five-factor model. *Psychological inquiry, 4(2)*, 100-104.

Clark, L. A., & Harrison, J. A. (2001). Assessment instruments. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (pp. 277-306). New York: Guildford.

Clark, L. A., Livesley, W. J., & Morey, L. (1997). Special feature: Personality disorder assessment: The challenge of construct validity. *Journal of Personality Disorders, 11*, 205-231.

Clark, L. A., Livesley, W. J., Schroeder, M. L., & Irish, S. L. (1996). Convergence of two systems for assessing specific traits of personality disorder. *Psychological Assessment, 8(3)*, 294-303.

Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology, 46*, 121-153.

Clarkin, J. F., Hull, J. W., Cantor, J., & Sanderson, C. (1993). Borderline personality disorder and personality traits: A comparison of SCID-II BPD and NEO-PI. *Psychological Assessment, 5(4)*, 472-476.

Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants. *Archives of General Psychiatry, 44*, 573-588.

Cloninger, C. R., & Svrakic, D. M. (1994). Differentiating normal and deviant personality by the seven-factor personality model. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 40-64). New York: Springer.

Coccaro, E. F. (2001). Biological and treatment correlates. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (2nd ed., pp. 124-135). New York: Guilford.

Cole, D. A. (2004). Taxometrics in psychopathology research: An introduction to some of the procedures ad related methodological issues. *Journal of Abnormal Psychology, 113(1)*, 3-9.

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual.* Champagne, IL: Institute for Personality and Ability Testing.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist – Revised. *Psychological Assessment, 9(1)*, 3-14.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78(1)*, 98-104.

Costa, P. T., & McCrae, R. R. (1990). Personality disorders and the five-factor model of personality. *Journal of Personality Disorders, 4(4)*, 362-371.

Costa, P. T., & McCrae, R. R. (1992a). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory professional manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1992b). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4(1)*, 5-13.

Costa, P. T., & Widiger, T. A. (1994). *Personality disorders and the five-factor model of personality* (1st ed.). Washington, DC: American Psychological Association.

Costa, P. T., & Widiger, T. A. (2002a). *Introduction: Personality disorders and the five-factor model of personality* (2nd ed.). In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 3-14). Washington, DC: American Psychological Association.

Costa, P. T., & Widiger, T. A. (2002b). *Personality disorders and the five-factor model of personality* (2nd ed.). Washington, DC: American Psychological Association.

Costa, P. T., Zonderman, A. B., McCrae, R. R., & Williams, R. B., (1985). Content and comprehensiveness in the MMPI: An item factor analysis in a normal adult sample. *Journal of Personality and Social Psychology, 48*, 925-933.

Davison, G. C., & Neale, J. M. (1996). *Abnormal psychology* (6th ed.). New York: John Wiley & Sons.

Depue, R. A., & Lezenweger, M. F. (2001). A neurobehavioral dimensional model. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (2nd ed., pp. 136-176). New York: Guilford.

Desinger, J. A. (1995). Exploring the factor structure of the Personality Assessment Inventory. *Assessment, 2(2)*, 173-179.

DeVellis, R. F. (1991). Scale development: Theory and applications. *Applied Social Research Methods Series (Volume 26)*. Newbury Park, CA: Sage.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Digman, J. M. (2002). Historical antecedents of the five-factor model. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 223-246). Washington, DC: American Psychological Association.

Dodd, B. G., & De Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement. Theory into practice* (Vol. 2, pp. 301-317). Norwood, NJ: Ablex.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13(2)*, 129-143.

Dolan-Sewell, R. T., Krueger, R. F., & Shea, M. T. (2001). Co-occurrence with syndrome disorders. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (2nd ed., pp. 84-104). New York: Guilford.

Douglas, K. S., Hart, S. D., & Kropp, P. R. (2001). Validity of the Personality Assessment Inventory for Forensic Assessments. *International Journal of Offender Therapy and Comparative Criminology, 45(2)*, 183-197.

Downey, R. G., Sinnett, E. R., Seeberger, W. (1998). The changing face of MMPI practice. *Psychological Reports, 83*, 1267-1272.

Dyce, J. A., & O'Connor, B. P. (1998). Personality disorders and the five-factor model: A test of facet-level predictions. *Journal of Personality Disorders, 12,* 31-45.

Dyce, J. A., O'Connor, B. P., Parkins, S., & Janzen, H. (1997). Correlational structure of the MCMI-III in personality disorder scales and comparisons with other data sets. *Journal of Personality Assessment, 69*, 568-582.

Edens, J. F., Hart, S. D., Johnson, D. W., Johnson, J. K., & Olver, M. E. (2000). Use of the Personality Assessment Inventory (PAI) to assess psychopathy in offender populations. *Psychological Assessment, 12*, 132-139.

Embretson, S. E., & Hershberger, S. L. (1999). Summary and future of psychometric methods in testing. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 242-). Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Eysenck, H. J. (1947). *Dimensions of personality*. London: Routledge & Kegan Paul.

Eysenck, H. J. (1952). *The scientific study of personality*. London: Routledge & Kegan Paul.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78(2)*, 350-365.

Funder, D. C. (2001). Personality. *Annual Review of Psychology, 52*, 197-221.

Grilo, C. M., & McGlashan, T. H. (2000). Convergent and discriminant validity of the DSM-IV axis II personality disorder criteria in adult outpatients with binge eating disorder. *Comprehensive Psychiatry, 41(3)*, 163-166.

Grilo, C. M., McGlashan, T. H., Morey, L. C., Gunderson, G., Skodol, A. E., Shea, M. T., et al. (2001). Internal consistence, intercriterion overlap and diagnostic efficiency of criteria

sets for DSM-IV schizotypal, borderline, avoidant and obsessive-compulsive personality disorders. *Acta Psychiatrica Scandinavica, 104(4)*, 264-272.

Grilo, C. M., Sanislow, C. A., Shea, M. T., Skodol, A. E., Stout, R. L., Gunderson, J. G., et al. (2005). Two-year prospective naturalistic study of remission from major depressive disorder as a function of personality disorder comorbidity. *Journal of Consulting and Clinical Psychology, 73(1)*, 78-85.

Hambleton, R. K. (1989). Principle and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Items: Instructions topics in Educational Measurement, Module 16*, 253-262.

Hare, R. D. (2006). Psychopathy: A clinical construct whose time has come. In C. R. Bartol & A. M. Bartol (Eds.), *Current perspectives in forensic psychology and criminal justice* (pp. 107-117). Thousand Oaks, CA: Sage.

Hare, R. D., & Hart, S. D. (1995). Commentary on antisocial personality disorder: The DSM-IV field trial. In W. J. Livesley (Ed.), *The DSM-IV personality disorders* (pp. 127-134). New York: Guilford.

Harper, T. J., Hakstian, A. R., & Hare, R. D. (1998). Factor structure of the Psychopathy Checklist. *Journal of Consulting and Clinical Psychology, 56(5)*, 741-747.

Haslam, N. (2003). The dimensional view of personality disorders: A review of taxometric evidence. *Clinical Psychology Review, 23*, 75-93.

Helmes, E., & Reddon, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin, 113*, 453-471.

Hershberger, S. L. (1999). Introduction to personality measurement. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 153-158). Mahwah, NJ: Lawrence Erlbaum Associates.

Hill, J., Fudge, H., Harrington, A. P., Pickles, A., & Rutter, M. (2000). Complementary approaches to the assessment of personality disorder. *British Journal of Psychiatry, 176*, 434-439.

Hirschfeld, R. M., Klerman, G. L., Clayton, P. F., Keller, M. B., McDonald-Scott, P., & Larkin, B. H. (1983). Assessing personality: Effects of depressive state on trait measurement. *American Journal of Psychiatry, 140*, 695-699.

Hirschfeld, R. M., Klerman, G. L., Lavori, P., Keller, M. B., Griffith, P., & Coryell, W. (1989). Premorbid personality assessments of first onset of major depression. *Archives of General Psychiatry, 46(4)*, 345-350.

Holaday, M., Smith, D. A., & Sherry, A. (2000). Sentence completion tests: A review of the literature and results of a survey of members of the society for personality assessment. *Journal of Personality Assessment, 74(3)*, 371-383.

Hurt, S. W., Clarkin, J. F., Widiger, T., Fyer, M., Sullivan, T., Stone, M., et al. (1990). Evaluation of DSM-III decision rules for case detection using joint conditional probability structures. *Journal of Personality Disorder, 4*, 121-130.

Jablensky, A. (2002). The classification of personality disorders: Critical review and need for rethinking. *Psychopathology, 35*, 112-116.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberg (Ed.), Current topics in clinical and community psychology (pp. 61-96). New York: Academic.

Jackson, K. M., & Trull, T. J. (2001). The factor structure of the Personality Assessment Inventory-Borderline features (PAI-BOR) scale in a nonclinical sample. Journal of Personality Disorders, 15(6), 536-545.

Johnson, J. H., Butcher, J. N., Null, C., & Johnson, K. N. (1984). Replicated item level factor analysis of the full MMPI. *Journal of Personality and Social Psychology, 47*, 105-114.

Johnson, J. G., Cohen, P., Kasen, S., Dkodol, A. E., Hamagami, F., & Brook, J. S. (2000). Age-related change in personality disorder trait levels between early adolescence and adulthood: A community-based longitudinal investigation. *Acta Psychiatrica Scandinavica, 102*, 265-275.

Kaplan, R. M., & Saccuzzo, D. P. (1993). *Psychological testing: Principles, applications, and issues* (3rd ed.). Belmont, CA: Brooks/Cole.

Kellogg, S. H., Ho, A., Bell, K., Schluger, R. P., McHugh, P. F., McClary, K. A., & Kreek, M. J. (2002). The Personality Assessment Inventory drug problems scale: A validity analysis. *Journal of Personality assessment, 79(1)*, 73-84.

Kernberg, O. F. (1996). Psychoanalytic theory of personality disorders. In J. F. Clarkin, & M. F. Lenzenweger (eds.), *Major theories of personality disorder* (pp. 106-140). New York: Guilford.

Kiehl, K. A., Smith, A. M., Mendrek, A., Forster, B. B., Hare, R. D., & Liddle, P. F. (2004). Temporal lobe abnormalities in semantic processing by criminal psychopaths as revealed

by functional magnetic resonance imaging. *Psychiatry Research: Neuroimaging, 130(1)*, 27-42.

Kiesler, D. J. (1986). The 1982 interpersonal circle: An analysis of DSM-III personality disorders. In T. Millon, & G. L. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 571-597). New York: Guildford.

Kiesler, D. J. (1996). *Contemporary interpersonal theory and research: Personality, psychopathology, and psychotherapy*. Oxford, England: John Wiley & Sons.

Kiesler, D. J. (2004). Using the impact message inventory-circumplex version to measure objective countertransference. Retrieved from: www.vcu.edu/sitar/IMI-CObjectiveCT.pdf; October 12, 2005.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E., & Surface, C. R. (1993). Enhancing the precision of the Mississippi scale for combat-related posttraumatic stress disorder: An application of item response theory. *Psychological Assessment, 5(4)*, 457-471.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25(2)*, 146-162.

Klonsky, E. D. (2004). Performance of Personality Assessment Inventory and Rorschach indices of schizophrenia in a public psychiatric hospital. *Psychological Services, 1(2)*, 107-110.

Kraemer, H. C., Noda, A., & O'Hara, R. (2003). Categorical versus dimensional approaches to diagnosis: Methodological challenges. *Journal of Psychiatric Research, 38(1)*, 17-25.

Kurtz, J. E., & Morey, L. C. (1998). Negativism in evaluative judgments of words among depressed outpatients with borderline personality disorder. *Journal of Personality Disorders, 12*, 351-361.

Kurtz, J. E., & Morey, L. C. (2001). Use of structured self-report assessment to diagnose borderline personality disorder during major depressive episodes. *Assessment, 8(3)*, 291-300.

Kwapil, T. R., Miller, M. B., Zinser, M. C., Chapman, J., & Chapman, L. J. (1997). Magical ideation and social anhedonia as predictors of psychosis proneness: A partial replication. *Journal of Abnormal Psychology, 106(3)*, 491-495.

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation*. Oxford, England: Ronald.

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice, 1(1)*, 71-83.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: Guilford Press.

Livesley, W. J. (2001). Conceptual and taxonomic issues. In W. J. Livesley (Ed.), *Handbook of personality disorders* (pp. 3-38). New York: Guilford Press.

Livesley, W. J., & Jackson, D. (in press). *Manual for the Dimensional Assessment of Personality Pathology-Basic Questionnaire*. Port Huron, MI: Sigma.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 51, 493-504.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Long, J. D., Harring, J. R., Brekke, J. S., Test, M. A. & Greenberg, J. (2007). Longitudinal construct validity of brief symptom inventory subscales in schizophrenia. *Psychological Assessment, 19(3)*, 298-308.

Loranger, A. W., Lezenweger, M. F., Gartner, A. F., Susman, V. L., Herzig, J., Zammit, G. K., et al. (1991). Trait-state artifacts and the diagnosis of personality disorders. *Archives of General psychiatry, 48*, 720-728.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23(2)*, 157-162.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubin, B., Wallis, R. R., & Paine, C. (1971). Patterns of psychological test usage in the United states: 1935-1969. *Professional Psychology, 2*, 70-74.

Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 129-152). Mahwah, NJ: Lawrence Erlbaum Associates.

Maremmani, I., Akiskal, H. S., Signoretta, S., Liguori, A., Perugi, G., & Cloninger, R. (2005). The relationship of Kraepelian affective temperaments (as measured by TEMPS-I) to the

tridimensional personality questionnaire (TPQ). *Journal of Affective Disorders, 85*, 17-27.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47(2)*, 149-174.

McCrae, R. R. (1994). A reformulation of Axis II: Personality and personality-related problems. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 303-309). Washington, DC: American Psychological Association.

McDevitt-Murphy, M. E. (2004). The utility of the PAI and the MMPI-2 for discriminating posttraumatic stress disorder, depression and social phobia in trauma-exposed college students [Abstract]. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 65*, 3171.

McDougall, W. (1932). Of the words of character and personality. *Character & Personality, 1*, 3-16.

McGlashan, T. H. (1986). Schizotypal personality disorder: Chestnut Lodge follow-up study: VI. Long-term follow-up perspectives. *Archives of General Psychiatry, 43(4)*, 329-334.

McMahon, R. C., Flynn, P. M., & Davidson, R. S. (1985). Stability of the personality and symptom scales of the Millon clinical Multiaxial Inventory. *Journal of Personality Assessment, 49(3)*, 231-234.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology, 113(1),* 39-43.

Mellsop, G., Varghese, F., Joshua, S., & Hicks, A. (1982). The reliability of axis II of DSM-III. *American Journal of Psychiatry, 139(10)*, 1360-1361.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Millon, T. (1981). *Disorders of personality: DSM-III, Axis II.* New York: Wiley.

Millon, T., & Davis, R. D. (1996). *Disorders of personality: DSM-IV and beyond.* New York: John Wiley & Sons.

Morey, L. C. (1988). A psychometric analysis of the DSM-III-R personality disorder criteria. *Journal of Personality Disorders, 2*, 109-124.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Morey, L. C., Skodol, A. E., Grilo, C. M., Sanislow, C. A., Zanarini, M. C., Shea, M. T., et al. (2004). Temporal coherence of criteria for four personality disorders. *Journal of Personality Disorders, 18(4)*, 394-398.

Morey, L. C., Waugh, M. H., & Blashfield, R. K. (1985). MMPI scales for DSM-III personality disorders: Their derivation and correlates. *Journal of Personality Assessment, 49*, 245-251.

Millon, T. (1969). *Modern psychopathology: A biosocial approach to maladaptive learning and functioning.* Philadelphia, PA: W. B. Saunders.

Millon, T. (1981). *Disorders of personality.* New York: John Wiley & Sons.

Millon, T. & Davis, R. O. (1996). *Disorders of personality: DSM-IV and beyond* (2nd ed.). Oxford, England: John Wiley & Sons.

Molenaar, I. W, & Sijtsma, K. (2000). *MSP5 for Windows, a program for Mokken scale analysis for polytomous items.* Groningen, The Netherlands: iec ProGAMMA.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Muraki, (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17(4)*, 351-363.

Nandakumar, Ratna, & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18(1)*, 41-68.

Newman, D. L., Moffitt, T. E., Caspi, A., Magdol, L., Silva, P. A., & Stanton, W. R. (1996). Psychiatric disorder in a birth cohort of young adults: Prevalence, co-morbidity, clinical significance, and new case incidence from age 11 to 21. *Journal of Consulting and Clinical Psychology, 64*, 552-562.

Newton-Howes, G., Tyrer, P., & Johnson, T. (2006). Personality disorder and the outcome of depression: Meta-analysis of published studies. *British Journal of Psychiatry, 188(1)*, 13-20.

O'Boyle, M., & Self, D. (1990). A comparison of two interviews for DSM-III-R personality disorders. *Psychiatry Research,* 32, 85-92.

O'Connor, B. P. (2002a). The search for dimensional structure difference between normality and abnormality: A statistical review of published data on personality and psychopathology. *Journal of Personality and Social Psychology, 83(4)*, 962-982.

O'Connor, B. P. (2002b). A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. *Assessment, 9(2),* 188-203.

O'Connor, B. P. (2005a). A search for consensus on the dimensional structure of personality disorders. *Journal of Clinical Psychology, 61(3),* 323-345.

O'Connor, B. P. (2005b). Graphical analyses of personality disorders in Five-Factor Model Space. *European Journal of Personality, 19,* 287-305.

O'Connor, B. P., & Dyce, J. A. (1998). A test of models of personality disorder configuration. *Journal of Abnormal Psychology, 107*, 3-16.

O'Connor, B. P., & Dyce, J. A. (2001). Rigid and extreme: A geometric representation of personality disorders in five-factor model space. *Journal of Personality and Social Psychology, 81(6),* 1119-1130.

O'Connor, B. P., & Dyce, J. A. (2002a). Tests of general and specific models of personality disorder configuration. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 223-246). Washington, DC: American Psychological Association.

O'Connor, B. P., & Dyce, J. A. (2002b). The search for dimensional structure differences between normality and abnormality: A statistical review of published data on personality and psychopathology. *Journal of Personality and Social Psychology, 83,* 962-982.

Oldham, J. M., Skodol, A. E., Kellman, H. D., Hyler, S. E., Rosnick, L., & Davies, M. (1992). Diagnosis of DSM-III-R personality disorders by two semistructured interviews: Patterns of comorbidity. *American Journal of Psychiatry, 149*, 213-220.

Ottosson, H., Grann, M., & Kullgren, G. (2000). Test-retest reliability of a self-report questionnaire for DSM-IV and ICD-10 personality disorders. *European Journal of Psychological Assessment, 16(1)*, 53-58.

Parchev, I. (2004). A visual guide to item response theory. Retrieved from: http://www.metheval.uni-jena.de/irt/VisualIRT.pdf; October 20, 2006.

Paris, J. (1987). Long-term follow-up of borderline patients in a general hospital. *Comprehensive Psychiatry, 28(6)*, 530-535.

Paris, J. (2003). Personality disorders over time: Precursors, course and outcome. Journal of Personality Disorders, 17(6), 479-488.

Paris, J., & Zweig-Frank, H. (2001). The 27-year follow-up of patient with borderline personality disorder. *Comprehensive Psychiatry, 42(6)*, 482-487.

Parker, J. D., Daleiden, E. L., & Simpson, C. A. (1999). Personality Assessment Inventory substance-use scales: Convergent and discriminant relations with the Addiction Severity Index in a residential chemical dependence treatment setting. *Psychological Assessment, 11(4)*, 507-513.

Patrick, C. J., Curtin, J. J., Tellegen, A. (2002). Development and validation of a brief form of the multidimensional personality questionnaire. *Psychological Assessment, 14(2)*, 150-163.

Perry, J. C. (1992). Problems and considerations in the valid assessment of personality disorders. *The American Journal of Psychiatry, 149(12)*, 1645-1653.

Pfhol, B. (1999). Axis I and Axis II: Comorbidity or confusion? In C. R. Cloninger (Ed.), *Personality and psychopathology* (pp. 83-98). Washington, DC: American Psychiatric Press.

Piedmont, R. L. (1998). *The revised NEO Personality Inventory: Clinical and research applications*. New York: Plenum.

Piersma, H. L. (1985). The Millon Clinical Multiaxial Inventory (MCMI) as a treatment outcome measure for psychiatric inpatients. *Journal of Clinical Psychology, 42(3)*, 493-499.

Piersma, H. L. (1987). The MCMI as a measure of DSM-III Axis II diagnoses: An empirical comparison. *Journal of Clinical Psychology, 43(5)*, 478-483.

Piersma, H. L. (1989). The MCMI as a treatment outcome measure for psychiatric patients. *Journal of Clinical Psychology, 45(1)*, 87-93.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorder: The use of the LEAD standard. *Psychological Assessment, 3*, 46-54.

Ping, C., Shuliang, D., Haijing, L., & Zhou, J. (2006). Item selection strategies of computerized adaptive testing based on graded response model [Abstract]. *Acta Psychologica Sinica, 38(3)*, 461-467.

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55(7)*, 787-796.

Piotrowski, C. (2000). How popular is the personality assessment inventory in practice and training? *Psychological Reports, 86*, 65-66.

Piotrowski, C., & Belter, R. W. (1999). Internship training in psychological assessment: Has managed care had an impact? *Assessment, 6*, 381-389.

Piotrowski, C., Sherry, D., & Keller, J. W. (1985). Psychodiagnostic test usage: A survey of the Society for Personality Assessment. *Journal of Personality Assessment, 49*, 115-119.

Rabin, L. A., Barr, W. B., Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33-65.

Ramsay, J. O. (2000). *TestGraf.* Retrieved January 2006, from http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2008). An IRT analysis of the personal optimism scale. *European Journal of Psychological Assessment, 24(1)*, 49-56.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21(1)*, 25-36.

Rehder, B. (2006). *Research*. Retrieved January, 2006, from http://www.psych.nyu.edu/rehder/.

Reich, J., Noyes, R., Coryell, W., & O'Gorman, T. W. (1986). The effect of state anxiety on personality measurement. *American Journal of Psychiatry, 143*, 760-763.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219-241). Mahwah, NJ: Lawrence Erlbaum Associates.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133-144.

Rizopoulos, D. (2006a). *Latent trait models under IRT*. Retrieved August, 2006, from http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:ltm

Rizopoulos, D. (2006b). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17(5)*, 1-25.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126(1)*, 3-25.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14(2)*, 187-207.

Rogers, R., Sewell, K. W., Ustad, K. L., Reinhardt, V., & Edwards, W. (1995). The Referral Decision Scale in a jail sample of disordered offenders. *Law and Human Behavior, 19*, 481-492.

Rogers, R., Ustad, K. L., Salekin, R. T. (1998). Convergent validity of the Personality Assessment Inventory: A study of emergency referrals in a correctional setting. *Assessment, 5(1)*, 3-12.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26(1)*, 42-56.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72(2)*, 282-307.

Rubio, V. J., Aguado, D., Hontangas, P. M., & Hernandez, J. M. (2007). Psychometric properties of an emotional adjustment measure. An application of the graded response model. *European Journal of Psychological Assessment, 23(1)*, 39-46.

Ruiz, M. A., Dickinson, K. A., & Pincus, A. L. (2002). Concurrent validity of the personality Assessment Inventory Alcohol Problems (ALC) Scale in a college student sample. *Assessment, 9(3)*, 261-270.

Salekin, R. T., Rogers, R., Ustad, K. L., Sewell, K. W. (1998). Psychopathy and recidivism among female inmates. Law & Human Behavior. Special Gender and the Law, 22(1), 109-128.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10(4)*, 345-359.

Schacht, T. E. (1993). How do I diagnose thee? Let me count the dimensions. *Psychological Inquiry, 4(2)*, 115-118.

Schinka, J. A. (1995). PAI profiles in alcohol-dependent patients. Journal of Personality *Assessment, 65(1)*, 35-51.

Schroeder, M. L., Wormworth, J. A., & Livesley, W. J. (1992). Dimensions of personality disorder and their relationships to the Big Five dimensions of personality. *Psychological Assessment, 4(1)*, 47-53.

Schroeder, M. L., Wormworth, J. A., & Livesley, W. J. (2002). Dimensions of personality disorder and the five-factor model of personality. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 149-160). Washington, DC: American Psychological Association.

Seivewright, H., Tyrer, P., & Johnson, T. (2002). Change in personality status in neurotic disorders. *Lancet, 359*, 2253-2254.

Shea, M. T., Stout, R. L., Yen, S., Pagano, M. E., Skodol, A. E., Morey, L. C., et al. (2004). Associations in the course of personality disorders and Axis I disorders over time. *Journal of Abnormal Psychology, 113(4)*, 499-508.

Shiner, R. L. (2005). A developmental perspective on personality disorders: Lessons from research on normal personality development in childhood and adolescence. *Journal of Personality Disorders, 19(2)*, 202-210.

Simms, L. J., Casillas, A., Clark, L. A., Watson, D., & Doebbeling, B. N. (2005). Psychometric evaluation of the restructured clinical scales of the MMPI-2. *Psychological Assessment, 17(3)*, 345-358.

Skinner, H. A. (1986). Construct validation approach to psychiatric classification. In T. Millon, & G. L. Klerman (Eds.). *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 307-330). New York: Guilford.

Skodol, A. E., Grilo, C. M., Pagano, M. E., Bender, D. S., Gunderson, J. G., Shea, M. T., et al. (2005). Effects of personality disorders on functioning and well-being in major depressive disorder. *Journal of Psychiatric Practice, 11(6)*, 363-368.

Skodol, A. E., Rosnick, L., Kellman, D., Oldham, J. M., & Hyler, S. E. (1988). Validating structured DSM-III-R personality disorder data. *The American Journal of Psychiatry, 145(10)*, 1297-1299.

Soderstrom, H. (2003). Psychopathy as a disorder of empathy. *European Child & Adolescent Psychiatry, 12*, 249-252.

Somwaru, D. P., & Ben-Porath, Y. S. (March, 1995). *Development and reliability of MMPI-2 based personality disorder scales.* Paper presented at the 30th Annual symposium on recent developments in the MMPI-2 and MMPI-A, St. Petersburg Beach, FL. (**I don't yet have this manuscript, but have emailed Ben-Porath to request a copy).

Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136(6)*, 815-817.

Steketee, G., Chambless, D. L., & Tran, G. Q. (2001). Effects of Axis I and II comorbidity on behavior therapy outcome for Obsessive-Compulsive disorder and Agoraphobia. *Comprehensive Psychiatry, 42(1)*, 76-86.

Stern, A. (1938). Borderline group of neuroses. *Psychoanalytic Quarterly, 7*, 467-489.

Strack, S., & Lorr, M. (1994). *Differentiating normal and abnormal personality*. New York: Springer.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: HarperCollins.

Tasca, G. A., Wood, J., Demindenko, N., Bissada, H. (2002). Using the PAI with an eating disordered population: Scale characteristics, factor structure, and differences among diagnostic groups. *Journal of Personality Assessment, 79(2)*, 337-356.

Tellegen, A. (1993). Folk concepts and psychological concepts of personality and personality disorder. *Psychological Inquiry, 4(2)*, 122-130.

Tellegen, A. (2006). *MPQ (Multidimensional Personality Questionnaire)*. Retrieved from http://www.upress.umn.edu/tests/mpq_overview.html, February 2006.

Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Trull, T. J. (1993). Temporal stability and validity of two personality disorder inventories. *Psychological Assessment, 5*, 11-18.

Trull, T. J. (1995). Borderline personality disorder features in nonclinical young adults: I. Identification and validation. *Psychological Assessment, 7(1)*, 33-41.

Trull, T. J. & Goodwin, A. H. (1993). Relationship between mood changes and the report of personality disorder symptoms. *Journal of Personality Assessment, 61(1)*, 99-111.

Trull, T. J., Useda, D., Conforti, K., & Doan, B. (1997). Borderline personality disorder features in nonclinical young adults: 1. Two-year outcome. *Journal of Abnormal Psychology, 106(2)*, 307-314.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 185-218). Mahwah, NJ: Erlbaum.

Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26(1)*, 54-60.

Westen, D. (1997). Divergences between clinical and research methods for assessing personality disorders: Implications for research and the evolution of Axis II. *The American Journal of psychiatry, 154*, 895-903.

Westen, D., & Arkowitz-Westen, L. (1998). Limitations of axis II in diagnosing personality pathology in clinical practice. *American Journal of Psychiatry, 155(12)*, 1767-1771.

Westen, D., & Shedler, J. (1999). Revising and assessing Axis II, Part II: Toward and empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156(2)*, 273-285.

White, L. J. (1996). Review of the Personality Assessment Inventory (PAI): A new psychological test for clinical and forensic assessment. *Australian Psychologist, 31(1)*, 38-39.

Widiger, T. A. (1998). Personality disorder dimensional models. In T. A. Widiger et al. (Eds.), *DSM-IV sourcebook: Vol. 4* (pp. 789-798). Washington, DC: American Psychiatric Association.

Widiger, T. A., & Anderson, K. G. (2003). Personality and depression in women. *Affective Disorders, 74*, 59-66.

Widiger, T. A., & Coker, L. A. (2002). Assessing personality disorders. In J. N. Butcher (Ed.), Clinical personality assessment: Practical approaches (pp. 407-434). New York: Oxford.

Widiger, T. A. & Costa, P. T. (1994). Personality and personality disorders. *Journal of Abnormal Psychology, 103*, 78-91.

Widiger, T. A., Frances, A. J. (1994). Toward a dimensional model for the personality disorders. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 19-39). Washington, DC: American Psychological Association.

Widiger, T. A., Frances, A. J., Harris, M., Jacobsberg, L. B., Fyer, M., & Manning, D. (1991). Comorbidity among Axis II disorders. In J. M. Oldham (Ed.), *Personality disorders: New perspective on diagnostic validity* (pp. 165-194). Washington, DC: American Psychiatric Association.

Widiger, T. A., & Samuel, D. B. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment, 3*, 278-287.

Widiger, T. A., & Trull, T. J. (1998). Performance characteristics of the DSM-III-R personality disorder criteria sets. In T. A. Widiger et al. (Eds.), *DSM-IV sourcebook: Vol. 4* (pp. 357-373). Washington, DC: American Psychiatric Association.

Wiggins, J. S. (1985). Symposium: Interpersonal circumplex models: 1948-1983: Commentary. *Journal of Personality Assessment, 49(6)*, 626-631.

Wiggins, J. S. (2003). *Paradigms of personality assessment.* New York: Guilford.

Wiggins, J. S., & Pincus, A. L. (1989). Conceptions of personality disorders and dimensions of personality. *Psychological Assessment, 1(4)*, 305-316.

Wiggins, J. S., & Pincus, A. L. (2002). Personality and the structure of personality disorders. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 103-124). Washington, DC: American Psychological Association.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Mooresville, IN: Scientific Software.

Winter, D. G., & Barenbaum, N. B. (1999). History of modern personality theory and research. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 3-27). New York: Guilford.

Wood, J. M., Garb, H. N., Lilienfeld, S. O., & Nezworski, M. T. (2002). Clinical assessment. *Annual Review of Psychology, 53*, 519-543.

Wright, B. D. (1967). *Sample free test calibration and person measurement*. Paper presented at the ETS Invitational Conference on Testing Problems. Retrieved January 23, 2005, from http://www.rasch.org/memo1.htm.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DL: Wide Range.

Yen, M., & Edwardson, S. R. (1999). Item- response theory approach in scale development. *Nursing Research, 48(4)*, 234-238.

Zanarini, M. C., Frankenburg, F. R., Vujanovic, A. A., Hennen, J., Reich, D. B., Silk, K. R. (2004). Axis II comorbidity of borderline personality disorder: Description of 6-year course and prediction to time-to-remission. *Acta Psychiatrica Scandinavica, 110(6)*, 416-420.

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry, 51*, 225-245.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99(3)*, 432-442.

Appendix A

PAI Scales

Table 1

PAI Scales

Full and subscales	Abbreviation	Number of items
Validity scales		
Inconsistency	ICN	20
Infrequence	INF	8
Negative Impression	NIM	9
Positive Impression	PIM	9
Clinical scales		
Somatic Complaints	SOM	24
Conversion	SOM-C	8
Somatization	SOM-S	8
Health Concerns	SOM-H	8
Anxiety	ANX	24
Affective	ANX-A	8
Physiological	ANX-P	8
Cognitive	ANX-C	8
Anxiety-Related Disorders	ARD	24
Obsessive-Compulsive	ARD-O	8
Phobias	ARD-P	8
Traumatic Stress	ARD-T	8
Depression	DEP	24
Cognitive	DEP-C	8
Affective	DEP-A	8
Physiological	DEP-P	8
Mania	MAN	24
Activity Level	MAN-A	8
Grandiosity	MAN-G	8
Irritability	MAN-I	8
Paranoia	PAR	24
Resentment	PAR-R	8
Hypervigilance	PAR-H	8
Persecution	PAR-P	8

PAI scales cont'd

Full and subscales	Abbreviation	Number of items
Schizophrenia	SCZ	24
Psychotic Experiences	SCZ-P	8
Social Detachment	SCZ-S	8
Thought Disorder	SCZ-T	8
Borderline Features	BOR	24
Affective Instability	BOR-A	6
Identity Problems	BOR-I	6
Negative Relationships	BOR-N	6
Self-Harm	BOR-S	6
Antisocial Features	ANT	24
Antisocial Behaviors	ANT-A	8
Egocentricity	ANT-E	8
Stimulus-Seeking	ANT-S	8
Alcohol Problems	ALC	12
Drug Problems	DRG	12
Treatment scales		
Aggression	AGG	18
Aggressive Attitude	AGG-A	6
Verbal Aggression	AGG-V	6
Physical Aggression	AGG-P	6
Suicidal Ideation	SUI	12
Stress	STR	8
Nonsupport	NON	8
Treatment Rejection	RXR	8
Interpersonal scales		
Dominance (/Submission)	DOM	12
Warmth (/Cold-rejecting)	WRM	12

Appendix B

PAI Items

PAI Items

Original Scale	PAI Items
NON	(1) My friends are available if I need them.
RXR	(2) I have some inner struggles that cause problems for me.
SOM-C	(3) My health condition has restricted my activities.
ANX-A	(4) I am so tense in some situations that I have great difficulty getting by.
ARD-O	(5) I have to do some things a certain way or I get nervous.
DEP-A	(6) Much of the time I'm sad for no real reason.
MAN-A	(7) Often I think and talk so quickly that other people cannot follow my train of thought.
PAR-H	(8) Most of the people I know can be trusted.
NIM	(9) Sometimes I cannot remember who I am.
SCZ-P	(10) I have some ideas that others think are strange.
ANT-A	(11) I was usually well-behaved at school.
SOM-H	(12) I've seen a lot of doctors over the years.
WRM	(13) I'm a very sociable person.
BOR-A	(14) My mood can shift quite suddenly.
ALC	(15) Sometimes I feel guilty about how much I drink.
DOM	(16) I'm a "take charge" type of person.
BOR-I	(17) My attitude about myself changes a lot.
AGG-V	(18) People would be surprised if I yelled at them.
BOR-N	(19) My relationships have been stormy.
SUI	(20) At times I wish I were dead.
AGG-P	(21) People are afraid of my temper.
DRG	(22) Sometimes I use drugs to feel better.
DRG	(23) I've tried just about every type of drug.
PIM	(24) Sometimes I let little things bother me too much.
ANX-C	(25) I often have trouble concentrating because I'm nervous.
ARD-P	(26) I often fear I might slip up and say something wrong.
DEP-C	(27) I feel that I've let everyone down.
MAN-G	(28) I have many brilliant ideas.
PAR-P	(29) Certain people go out of their way to bother me.
SCZ-S	(30) I just don't seem to relate to people very well.
ANT-E	(31) I've borrowed money knowing I wouldn't pay it back.
SOM-S	(32) Much of the time I don't feel well.
ANX-P	(33) I often feel jittery.
ARD-T	(34) I keep reliving something horrible that happened to me.
DEP-P	(35) I hardly have any energy.
MAN-I	(36) I can be very demanding when I want things done quickly.
PAR-R	(37) People usually treat me pretty fairly.

SCZ-T	(38) My thinking has become confused.
ANT-S	(39) I get a kick out of doing dangerous things.
INF	(40) My favorite poet is Raymond Kertezc.
NON	(41) I like being around my family.
RXR	(42) I need to make some important changes in my life.
SOM-C	(43) I've had illnesses that my doctors could not explain.
ANX-A	(44) I can't do some things well because of nervousness.
ARD-O	(45) I have impulses that I fight to keep under control.
DEP-A	(46) I've forgotten what it's like to feel happy.
MAN-A	(47) I take on so many commitments that I can't keep up.
PAR-H	(48) I have been alert to the possibility that people will be unfaithful.
NIM	(49) I have visions in which I see myself forced to commit crimes.
SCZ-P	(50) Other people sometimes put thoughts into my head.
ANT-A	(51) I've deliberately damaged someone's property.
SOM-H	(52) My health concerns are very complicated.
WRM	(53) It's easy for me to make new friends.
BOR-A	(54) My moods get quite intense.
ALC	(55) I have trouble controlling my use of alcohol.
DOM	(56) I'm a natural leader.
BOR-I	(57) Sometimes I feel terribly empty inside.
AGG-V	(58) I tell people off when they deserve it.
BOR-N	(59) I want to let certain people know how much they've hurt me.
SUI	(60) I've thought about ways to kill myself.
AGG-P	(61) Sometimes my temper explodes and I completely lose control.
DRG	(62) People have told me that I have a drug problem.
DRG	(63) I never use drugs to help me cope with the world.
PIM	(64) Sometimes I'll avoid someone I really don't like.
ANX-C	(65) It's often hard for me to enjoy myself because I am worrying about things.
ARD-P	(66) I have exaggerated fears.
DEP-C	(67) Sometimes I think I'm worthless.
MAN-G	(68) I have some very special talents that few others have.
PAR-P	(69) Some people do things to make me look bad.
SCZ-S	(70) I don't have much to say to anyone.
ANT-E	(71) I'll take advantage of others if they leave themselves open to it.
SOM-S	(72) I suffer from a lot of pain.
ANX-P	(73) I worry so much that at times I feel like I am going to faint.
ARD-T	(74) Thoughts about my past often bother me while I'm thinking about something else.
DEP-P	(75) I have no trouble falling asleep.
MAN-I	(76) I get quite irritated if people try to keep me from accomplishing my goals.
PAR-R	(77) I seem to have as much luck in life as others.
SCZ-T	(78) My thoughts get scrambled sometimes.

ANT-S	(79) I do a lot of wild things just for the thrill of it.
INF	(80) Sometimes I get ads in the mail that I don't really want.
NON	(81) If I'm having problems, I have people I can talk to.
RXR	(82) I need to change some things about myself, even if it hurts.
SOM-C	(83) I've had numbness in parts of my body that I can't explain.
ANX-A	(84) Sometimes I am afraid for no reason.
ARD-O	(85) It bothers me when things are out of place.
DEP-A	(86) Everything seems like a big effort.
MAN-A	(87) Recently I've had much more energy than usual.
PAR-H	(88) Most people have good intentions.
NIM	(89) Since the day I was born, I was destined to be unhappy.
SCZ-P	(90) Sometimes it seems that my thoughts are broadcast so that others can hear them.
ANT-A	(91) I've done some things that weren't exactly legal.
SOM-H	(92) It's a struggle for me to get things done with the medical problems I have.
WRM	(93) I like to meet new people.
BOR-A	(94) My mood is very steady.
ALC	(95) There have been times when I've had to cut down on my drinking.
DOM	(96) I would be good at a job where I tell others what to do.
BOR-I	(97) I worry a lot about other people leaving me.
AGG-V	(98) When I get mad at other drivers on the road, I let them know it.
BOR-N	(99) People once close to me have let me down.
SUI	(100) I've made plans about how to kill myself.
AGG-P	(101) Sometimes I'm very violent.
DRG	(102) My drug use has caused me financial strain.
DRG	(103) I've never had problems at work because of drugs.
PIM	(104) I sometimes complain too much.
ANX-C	(105) I'm often so worried and nervous that I can barely stand it.
ARD-P	(106) I get very nervous when I have to do something in front of others.
DEP-C	(107) I don't feel like trying anymore.
MAN-G	(108) My plans will make me famous someday.
PAR-P	(109) People around me are faithful to me.
SCZ-S	(110) I'm a loner.
ANT-E	(111) I'll do most things if the price is right.
SOM-S	(112) I am in good health.
ANX-P	(113) Sometimes I feel dizzy when I've been under a lot of pressure.
ARD-T	(114) I've been troubled by memories of a bad experience for a long time.
DEP-P	(115) I rarely have trouble sleeping.
MAN-I	(116) Sometimes I get upset because others don't understand my plans.
PAR-R	(117) I've given a lot, but I haven't gotten much in return.
SCZ-T	(118) Sometimes I have trouble keeping different thoughts separate.
ANT-S	(119) My Behaviour is pretty wild at times.

INF	(120) My favorite sports event on television is the high jump.
NON	(121) I spend most of my time alone.
RXR	(122) I need some help to deal with important problems.
SOM-C	(123) I've had episodes of double vision or blurred vision.
ANX-A	(124) I'm not the kind of person who panics easily.
ARD-O	(125) I can relax even if my home is a mess.
DEP-A	(126) Nothing seems to give me much pleasure.
MAN-A	(127) At times my thoughts move very quickly.
PAR-H	(128) I usually assume people are telling the truth.
NIM	(129) I think I have three or four completely different personalities inside of me.
SCZ-P	(130) Others can read my thoughts.
ANT-A	(131) I used to lie a lot to get out of tight situations.
SOM-H	(132) My medical problems always seem to be hard to treat.
WRM	(133) I am a warm person.
BOR-A	(134) I have little control over my anger.
ALC	(135) My drinking seems to cause problems in my relationships with others.
DOM	(136) I have trouble standing up for myself.
BOR-I	(137) I often wonder what I should do with my life.
AGG-V	(138) I'm not afraid to yell at someone to get my point across.
BOR-N	(139) I rarely feel very lonely.
SUI	(140) I've recently been thinking about suicide.
AGG-P	(141) Sometimes I smash things when I'm upset.
DRG	(142) I never use illegal drugs.
BOR-S	(143) I sometimes do things so impulsively that I get into trouble.
PIM	(144) Sometimes I'm too impatient.
ANX-C	(145) My friends say I worry too much.
ARD-P	(146) I'm not easily frightened.
DEP-C	(147) I can't seem to concentrate very well.
MAN-G	(148) I have accomplished some remarkable things.
PAR-P	(149) Some people try to keep me from getting ahead.
SCZ-S	(150) I don't feel close to anyone.
ANT-E	(151) I can talk my way out of just about anything.
SOM-S	(152) I seldom have complaints about how I feel physically.
ANX-P	(153) I can often feel my heart pounding.
ARD-T	(154) I can't seem to get over something from my past.
DEP-P	(155) I've been moving more slowly than usual.
MAN-I	(156) I have great plans and it irritates me that people try to interfere.
PAR-R	(157) People don't appreciate what I've done for them.
SCZ-T	(158) Sometimes it feels as if somebody is blocking my thoughts.
ANT-S	(159) If I get tired of a place, I just pick up and leave.
INF	(160) Most people would rather win than lose.

NON	(161) Most people I'm close to are very supportive.
RXR	(162) I'm curious why I behave the way I do.
SOM-C	(163) There have been times when my eyesight got worse and then better again.
ANX-A	(164) I am a very calm and relaxed person.
ARD-O	(165) People say that I'm a perfectionist.
DEP-A	(166) I've lost interest in things I used to enjoy.
MAN-A	(167) My friends can't keep up with my social activities.
PAR-H	(168) People generally hide their real motives.
NIM	(169) People don't understand how much I suffer.
SCZ-P	(170) I've heard voices that no one else could hear.
ANT-A	(171) I like to see how much I can get away with.
SOM-H	(172) I've had only the usual health problems that most people have.
WRM	(173) It takes me a while to warm up to people.
BOR-A	(174) I've always been a pretty happy person.
ALC	(175) Drinking helps me get along in social situations.
DOM	(176) I feel best in situations where I am the leader.
BOR-I	(177) I can't handle separation from those close to me very well.
AGG-V	(178) I always avoid arguments if I can.
BOR-N	(179) I've made some real mistakes in the people I've picked as friends.
SUI	(180) I have thought about suicide for a long time.
AGG-P	(181) I've threatened to hurt people.
DRG	(182) I've used prescription drugs to get high.
BOR-S	(183) When I'm upset, I typically do something to hurt myself.
PIM	(184) I don't take criticism very well.
ANX-C	(185) I don't worry about things any more than most people.
ARD-P	(186) I don't mind driving on freeways.
DEP-C	(187) No matter what I do, nothing works.
MAN-G	(188) I think I have the answers to some very important questions.
PAR-P	(189) There are people who want to hurt me.
SCZ-S	(190) I enjoy the company of other people.
ANT-E	(191) I don't like being tied to one person.
SOM-S	(192) I have a bad back.
ANX-P	(193) It's easy for me to relax.
ARD-T	(194) I have had some horrible experiences that make me feel guilty.
DEP-P	(195) I often wake up very early in the morning and can't get back to sleep.
MAN-I	(196) It bothers me when other people are too slow to understand my ideas.
PAR-R	(197) Usually I've gotten credit for what I've done.
SCZ-T	(198) My thoughts tend to quickly shift around to different things.
ANT-S	(199) The idea of "settling down" has never appealed to me.
INF	(200) My favorite hobbies are archery and stamp-collecting.
NON	(201) People I know care about me.

RXR	(202) I'm comfortable with myself the way I am.
SOM-C	(203) I've had episodes when I've lost the feeling in my hands.
ANX-A	(204) I often feel as if something terrible is about to happen.
ARD-O	(205) I'm usually aware of objects that have a lot of germs.
DEP-A	(206) I have no interest in life.
MAN-A	(207) I feel like I need to keep active and not rest.
PAR-H	(208) People think I'm too suspicious.
NIM	(209) Every once in a while I totally lost my memory.
SCZ-P	(210) There are people who try to control my thoughts.
ANT-A	(211) I was never expelled or suspended from school when I was young.
SOM-H	(212) I've had some unusual diseases and illnesses.
WRM	(213) It takes a while for people to get to know me.
BOR-A	(214) I've had times when I was so mad I couldn't do enough to express all my anger.
ALC	(215) Some people around me think I drink too much alcohol.
DOM	(216) I prefer to let others make decisions.
BOR-I	(217) I don't get bored very easily.
AGG-V	(218) I don't like raising my voice.
BOR-N	(219) Once someone is my friend, we stay friends.
SUI	(220) Death would be a relief.
AGG-P	(221) I've never started a physical fight as an adult.
DRG	(222) My drug use is out of control.
BOR-S	(223) I'm too impulsive for my own good.
PIM	(224) Sometimes I put things off until the last minute.
ANX-C	(225) I don't worry about things that I can't control.
ARD-P	(226) I don't mind heights.
DEP-C	(227) I think good things will happen to me in the future.
MAN-G	(228) I think I would be a good comedian.
PAR-P	(229) People seldom treat me badly on purpose.
SCZ-S	(230) I like to be around other people if I can.
ANT-E	(231) I don't like to stay in a relationship very long.
SOM-S	(232) I have a weak stomach.
ANX-P	(233) When I'm under a lot of pressure, I sometimes have trouble breathing.
ARD-T	(234) I keep having nightmares about my past.
DEP-P	(235) I have a good appetite.
MAN-I	(236) I have no patience with people who try to hold me back.
PAR-R	(237) People who are successful generally earned their success.
SCZ-T	(238) Sometimes I wonder if my thoughts are being taken away.
ANT-S	(239) I like to drive fast.
INF	(240) I don't like to have to buy things that are overpriced.
NON	(241) In my family, we argue more than we talk.
RXR	(242) Many of my problems are my own doing.

SOM-C	(243) I've had times when my legs became so weak that I couldn't walk.
ANX-A	(244) I seldom feel anxious or tense.
ARD-O	(245) People see me as a person who pays a lot of attention to detail.
DEP-A	(246) Lately I've been happy much of the time.
MAN-A	(247) Recently I have needed less sleep than usual.
PAR-H	(248) Things are rarely as they seem on the surface.
NIM	(249) Sometimes my vision is only in black and white.
SCZ-P	(250) I have a sixth sense that tells me what is going to happen.
ANT-A	(251) I've never been in trouble with the law.
SOM-H	(252) For my age, my health is pretty good.
WRM	(253) I try to include people who seem left out.
ALC	(254) Sometimes I have an alcoholic drink first thing in the morning.
ALC	(255) My drinking has caused me problems at home.
DOM	(256) I say what's on my mind.
DOM	(257) I usually do what other people tell me to do.
AGG-A	(258) I have a bad temper.
AGG-A	(259) It takes a lot to make me angry.
SUI	(260) I've thought about what I would say in a suicide note.
SUI	(261) I can't think of reasons to go on living.
DRG	(262) I've had health problems because of my drug use.
BOR-S	(263) I spend money too easily.
PIM	(264) I sometimes make promises I can't keep.
ANX-C	(265) I usually worry about things more than I should.
ARD-P	(266) I will not ride in airplanes.
DEP-C	(267) I have something worthwhile to contribute.
MAN-G	(268) Lately I feel so confident that I think I can accomplish anything.
PAR-P	(269) People have had it in for me.
SCZ-S	(270) I make friends easily.
ANT-E	(271) I look after myself first; let others take care of themselves.
SOM-S	(272) I get more headaches than most people.
ANX-P	(273) I get sweaty hands often.
ARD-T	(274) Since I had a very bad experience, I am no longer interested in some things that I used to enjoy.
DEP-P	(275) I often wake up in the middle of the night.
MAN-I	(276) At times I am very touchy and easily annoyed.
PAR-R	(277) I'm not the type of person to hold a grudge.
SCZ-T	(278) Thoughts in my head suddenly disappear.
ANT-S	(279) I'm not a person who turns down a dare.
INF	(280) Most people look forward to a trip to the dentist.
NON	(281) I spend little time with my family.
RXR	(282) I can solve my problems by myself.
SOM-C	(283) At times parts of my body have been paralyzed.

ANX-A	(284) I am easily startled.
ARD-O	(285) I keep myself under tight control.
DEP-A	(286) I'm almost always a happy and positive person.
MAN-A	(287) I hardly ever buy things on impulse.
PAR-H	(288) People have to earn my trust.
NIM	(289) I don't have any good memories from my childhood.
SCZ-P	(290) I don't believe that there are people who can read minds.
ANT-A	(291) I've never taken money or property that wasn't mine.
SOM-H	(292) I like to talk with people about their medical problems.
WRM	(293) I'm an affectionate person.
ALC	(294) I never drive when I've been drinking.
ALC	(295) I hardly ever drink alcohol.
DOM	(296) People listen to my opinions.
DOM	(297) If I get poor service from a business, I let the manager know about it.
AGG-A	(298) My temper never gets me into trouble.
AGG-A	(299) My anger never gets out of control.
SUI	(300) I've thought about how others would react if I killed myself.
SUI	(301) I have a lot to live for.
DRG	(302) My best friends are those I use drugs with.
BOR-S	(303) I'm a reckless person.
PIM	(304) There have been times when I could have been more thoughtful that I was.
ANX-C	(305) Sometimes I get so nervous that I'm afraid I'm going to die.
ARD-P	(306) I don't mind traveling in a bus or train.
DEP-C	(307) I'm pretty successful at what I do.
MAN-G	(308)I could never imagine myself being famous.
PAR-P	(309) I'm a target of a conspiracy.
SCZ-S	(310) I keep in touch with my friends.
ANT-E	(311) When I make a promise, I really don't need to keep it.
SOM-S	(312) I frequently have diarrhea.
ANX-P	(313) I have very steady hands.
ARD-T	(314) I avoid certain things that bring back bad memories.
DEP-P	(315) I have little interest in sex.
MAN-I	(316) I have little patience with those who disagree with my plans.
PAR-R	(317) Being helpful to other people pays off in the end.
SCZ-T	(318) I can concentrate now as well as I ever could.
ANT-S	(319) I never take risks if I can avoid it.
INF	(320) In my free time I might read, watch TV, or just relax.
STR	(321) I have a lot of money problems.
STR	(322) My life is very unpredictable.
STR	(323) There have been many changes in my life recently.
STR	(324) There isn't much stability at home.

STR	(325) Things are not going well in my family.
STR	(326) I'm happy with my job situation.
STR	(327) I worry about having enough money to get by.
STR	(328) My relationship with my spouse or partner is not going well.
NIM	(329) I have severe psychological problems that began very suddenly.
WRM	(330) I'm a sympathetic person.
WRM	(331) Close relationships are important to me.
WRM	(332) I'm very impatient with people.
WRM	(333) I have more friends than most people I know.
ALC	(334) My drinking has never gotten me into trouble.
ALC	(335) My drinking has caused problems with my work.
DOM	(336) I don't like letting people know when I disagree with them.
DOM	(337) I'm a very independent person.
AGG-A	(338) When I get mad, it's hard for me to calm down.
AGG-A	(339) People think I'm aggressive.
SUI	(340) I'm considering suicide.
SUI	(341) Things have never been so bad that I thought about suicide.
DRG	(342) My drug use has never caused problems with my family or friends.
BOR-S	(343) I'm careful about how I spend my money.
PIM	(344) I rarely get in a bad mood.

Appendix C

PAR, AVD, SZD, and SZT Mean Prototypicality Ratings ≥ 3.0

Table 6

PAR, AVD, SZD, and SZT Mean Prototypicality Ratings ≥ 3.0

PAR		AVD		SZD		SZT	
Item	Proto	Item	Proto	Item	Proto	Item	Proto
11	-4.17	4	4.33	13	-5.00	10	5.00
8	-5.00	13	-4.83	16	-4.00	13	-4.83
29	5.00	16	-4.00	41	-4.83	30	4.50
37	-5.00	18	4.50	53	-4.83	53	-4.33
48	5.00	21	-4.50	54	-4.17	78	4.17
53	-4.50	26	5.00	56	-4.50	93	-4.17
69	5.00	30	4.33	64	4.17	110	4.33
81	-4.00	53	-5.00	70	4.17	121	4.33
88	-5.00	56	-4.67	93	-5.00	130	4.67
99	5.00	58	-4.17	96	-4.33	133	-4.33
109	-5.00	64	4.00	97	-4.83	150	4.33
128	-5.00	65	4.00	110	5.00	158	4.33
149	5.00	93	-4.50	121	5.00	167	-4.00
150	4.17	96	-4.67	133	-4.83	170	4.33
161	-4.17	106	5.00	139	4.33	173	4.67
168	5.00	110	4.33	150	5.00	190	-4.33
173	4.50	121	4.17	167	-5.00	213	4.00
179	4.33	138	-4.33	173	4.00	230	-4.33
189	4.83	167	-4.17	176	-4.33	250	4.83
208	4.83	173	4.33	177	-4.83	270	-4.50
229	-4.83	176	-4.83	190	-5.00	288	4.33
269	5.00	178	4.67	213	4.83	290	-5.00
270	-4.00	184	5.00	238	-4.67	293	-4.00
277	-5.00	213	5.00	230	-5.00	333	-4.33
288	5.00	216	4.50	253	-4.50		
309	4.50	228	-4.50	270	-4.83		
333	-4.00	244	-4.50	282	4.00		
		256	-4.33	292	-4.00		
		270	-4.83	293	-5.00		
		297	-4.17	310	-5.00		
		333	-4.83	315	5.00		
		336	4.67	330	-4.83		
				331	-5.00		
				333	-5.00		
				337	4.50		

Appendix D

PAR

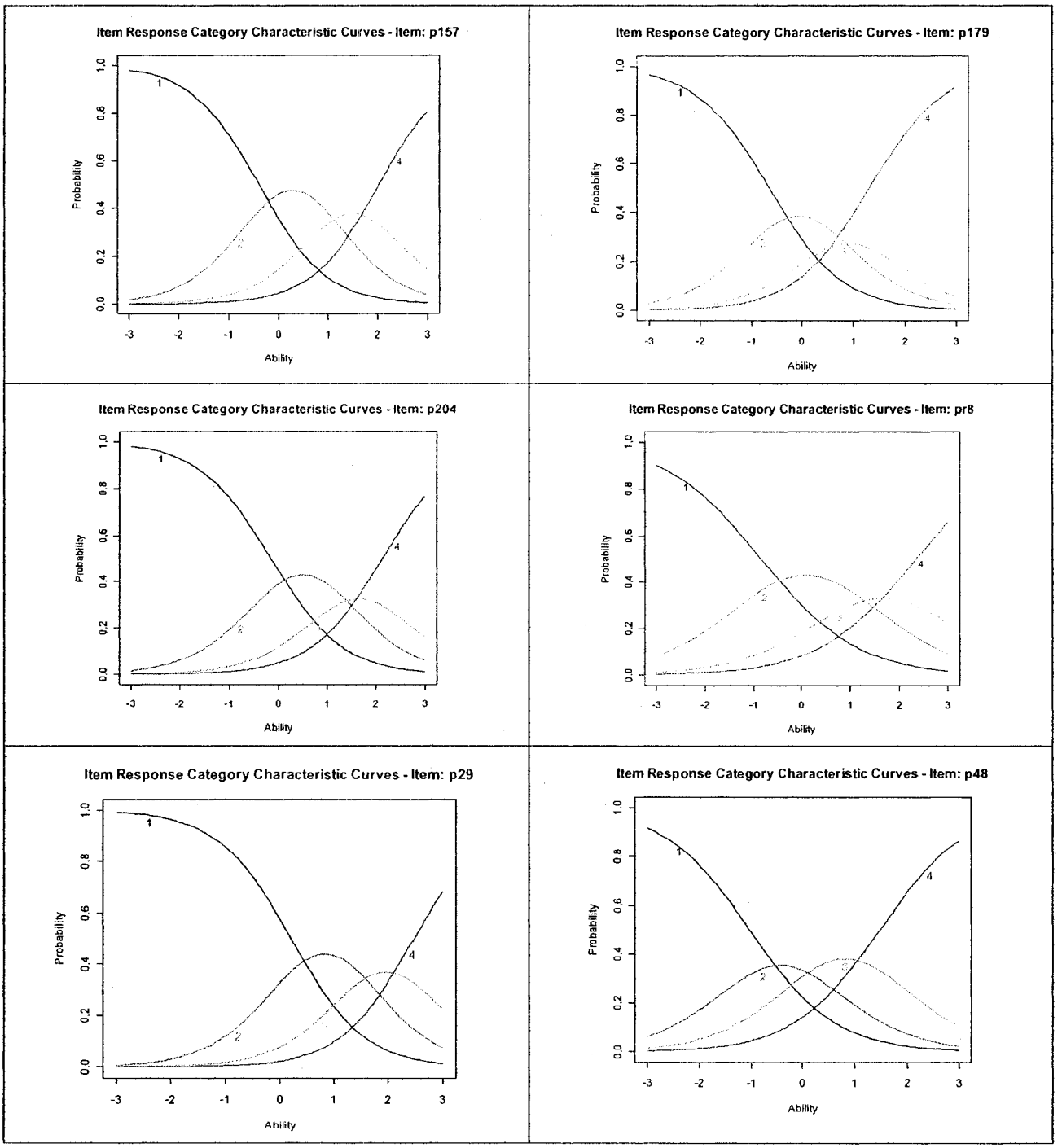

Figure 5. Item Response Category Characteristic Curves (CCC) for PAR Items (1-6)

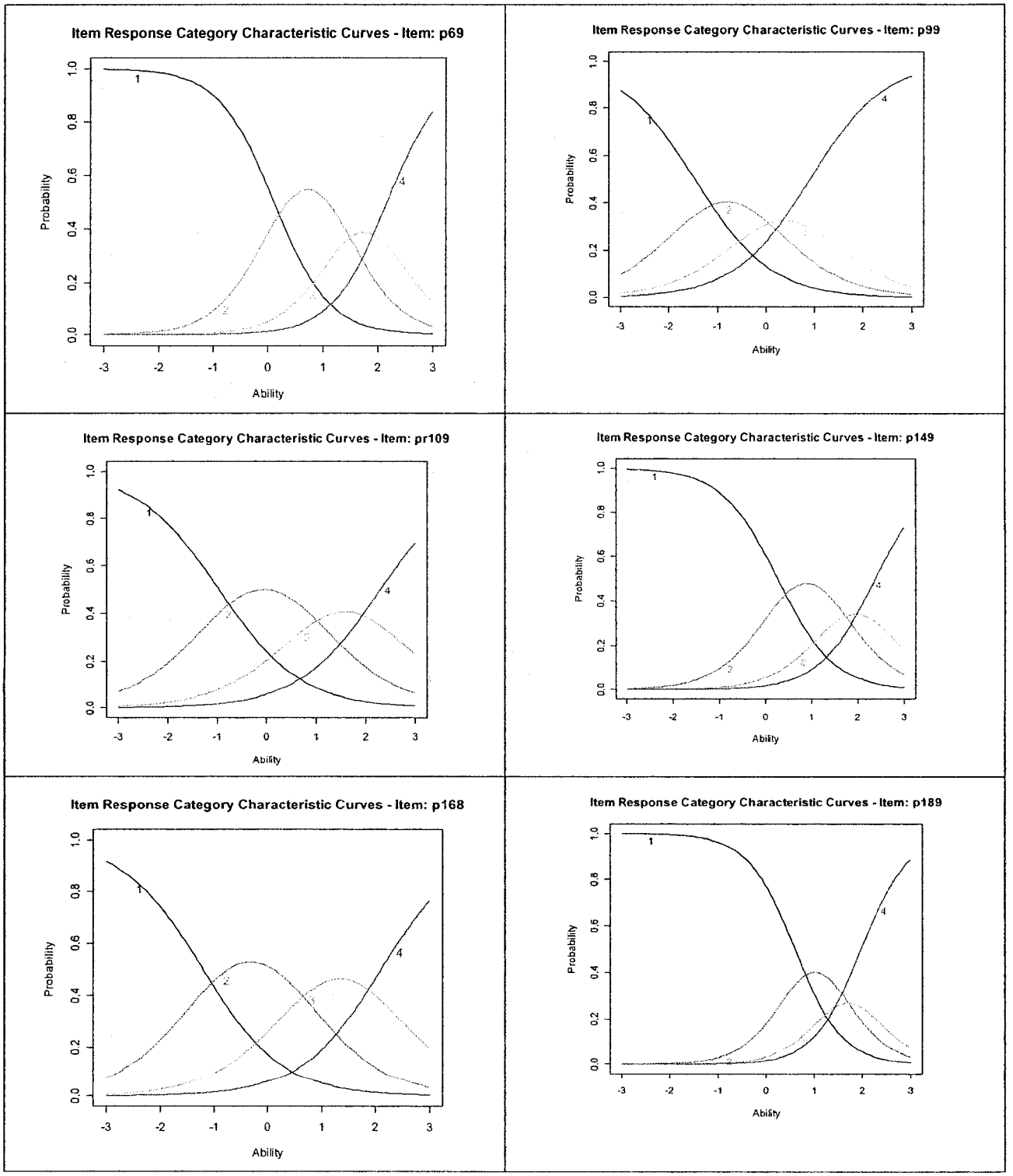

Figure 5 cont'd. Item Response Category Characteristic Curves (CCC) for PAR Items (7-12)

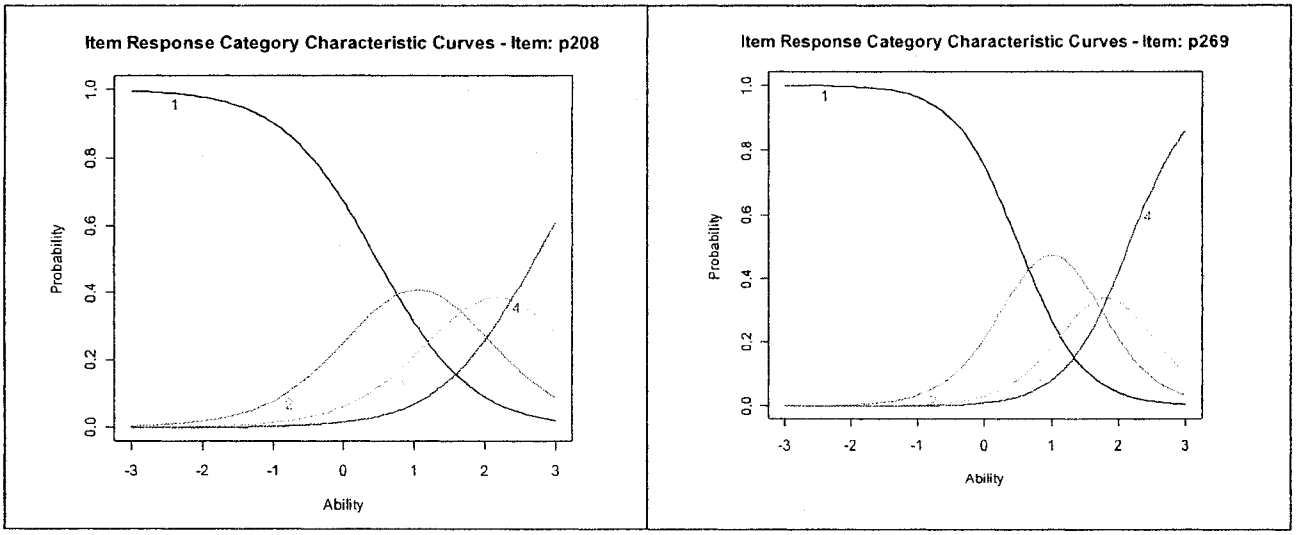

Figure 5 cont'd. Item Response Category Characteristic Curves (CCC) for PAR Items (13-14)

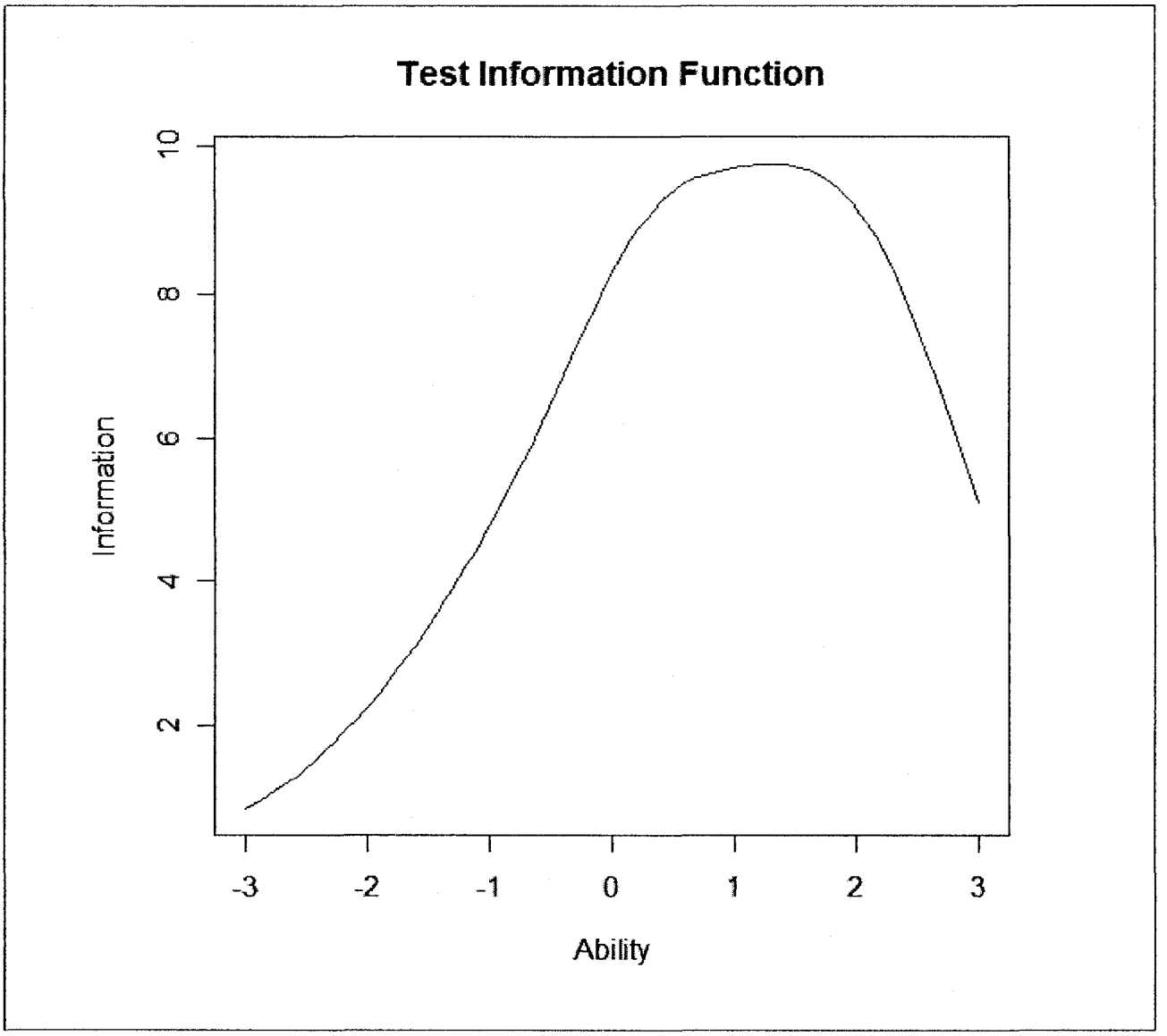

Figure 6. Test Information Function for the PAR scale.

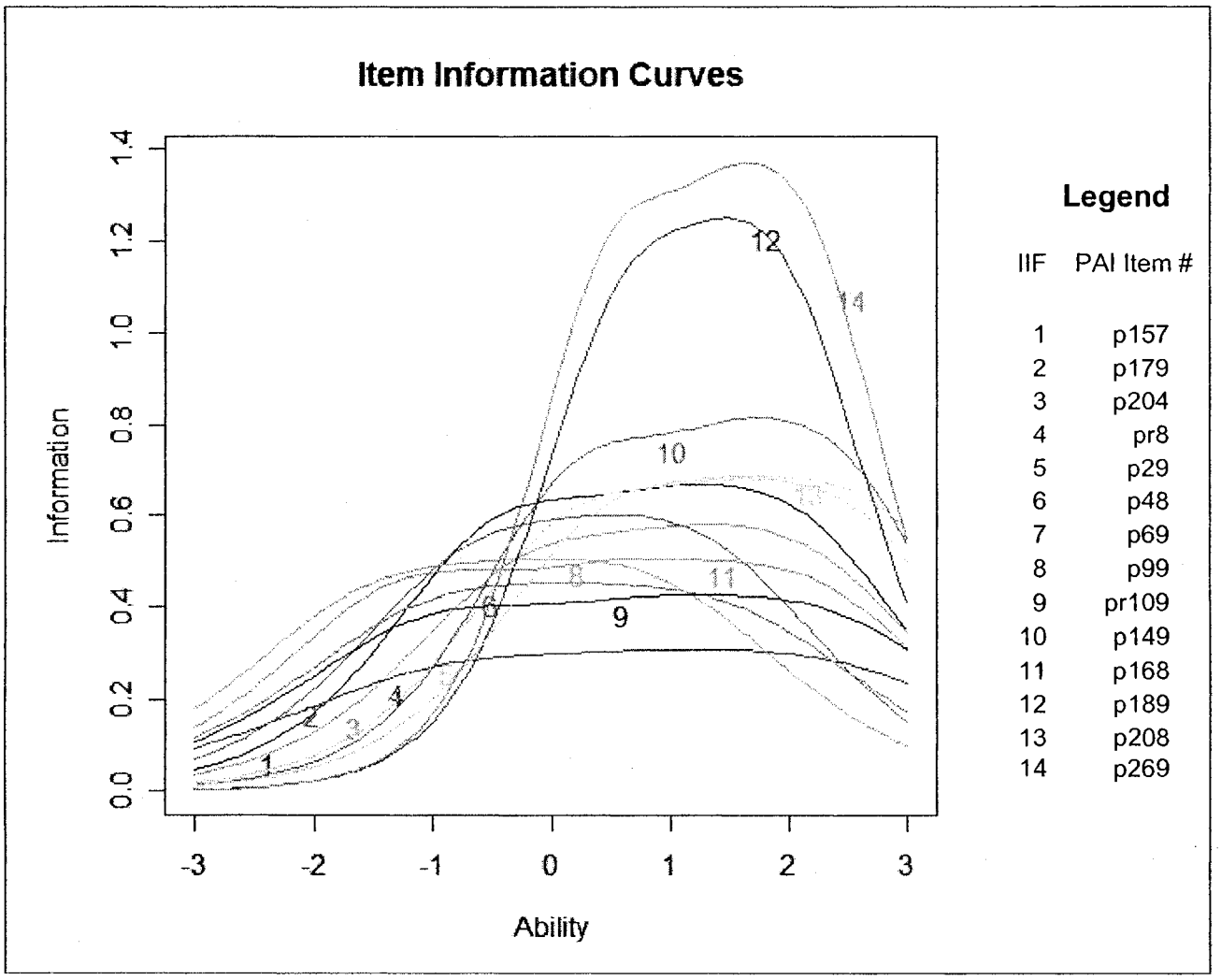

Figure 7. Item Information Functions for the PAR scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 8.1

PAR: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p157	-0.40	0.98	2.04	1.49	.40
p179	-0.65	0.51	1.31	1.40	.41
p204	-0.16	1.17	2.14	1.38	.40
pr8	-0.83	1.00	2.35	1.01	.36
p29	0.20	1.46	2.48	1.50	.38
p48	-1.03	0.18	1.48	1.22	.37
p69	0.10	1.34	2.17	1.97	.45
p99	-1.47	-0.14	0.91	1.28	.39
pr109	-0.97	0.86	2.30	1.20	.37
p149	0.26	1.53	2.39	1.64	.40
p168	-1.20	0.58	2.11	1.32	.37
p189	0.60	1.44	1.98	2.02	.43
p208	0.47	1.63	2.71	1.51	.39
p269	0.52	1.49	2.15	2.13	.38
Mean	-0.33	1.0	2.04	1.51	*(39)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 8.2

PAR: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	88.08
-2 to +2	67.43
-3 to 0	26.57
0 to +3	61.51
-3 to -2	3.4
-2 to -1	7.96
-1 to 0	15.21
0 to +1	21.68
+1 to +2	22.57
+2 to +3	17.25

Note. [1] = PAR trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 8.3

PAR: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p269	10.18
p69	10.07
p189	8.82
p149	7.89
p157	7.28
p29	7.16
p208	7.14
p168	6.95
p204	6.34
p179	6.06
pr109	6.04
p99	5.83
p48	5.55
pr8	4.68

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix E

SZD

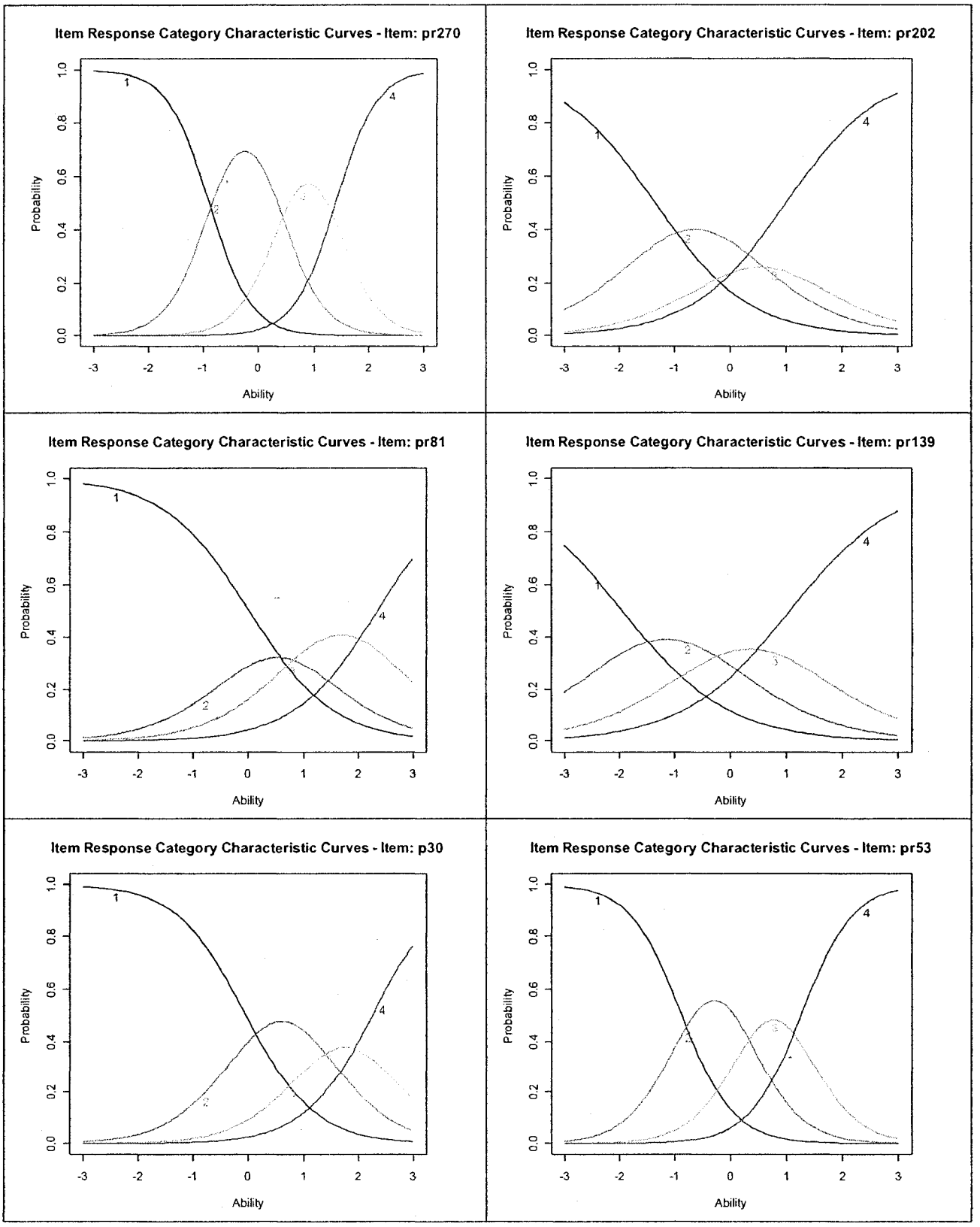

Figure 8. Item Response Category Characteristic Curves (CCC) for SZD Items (1-6)

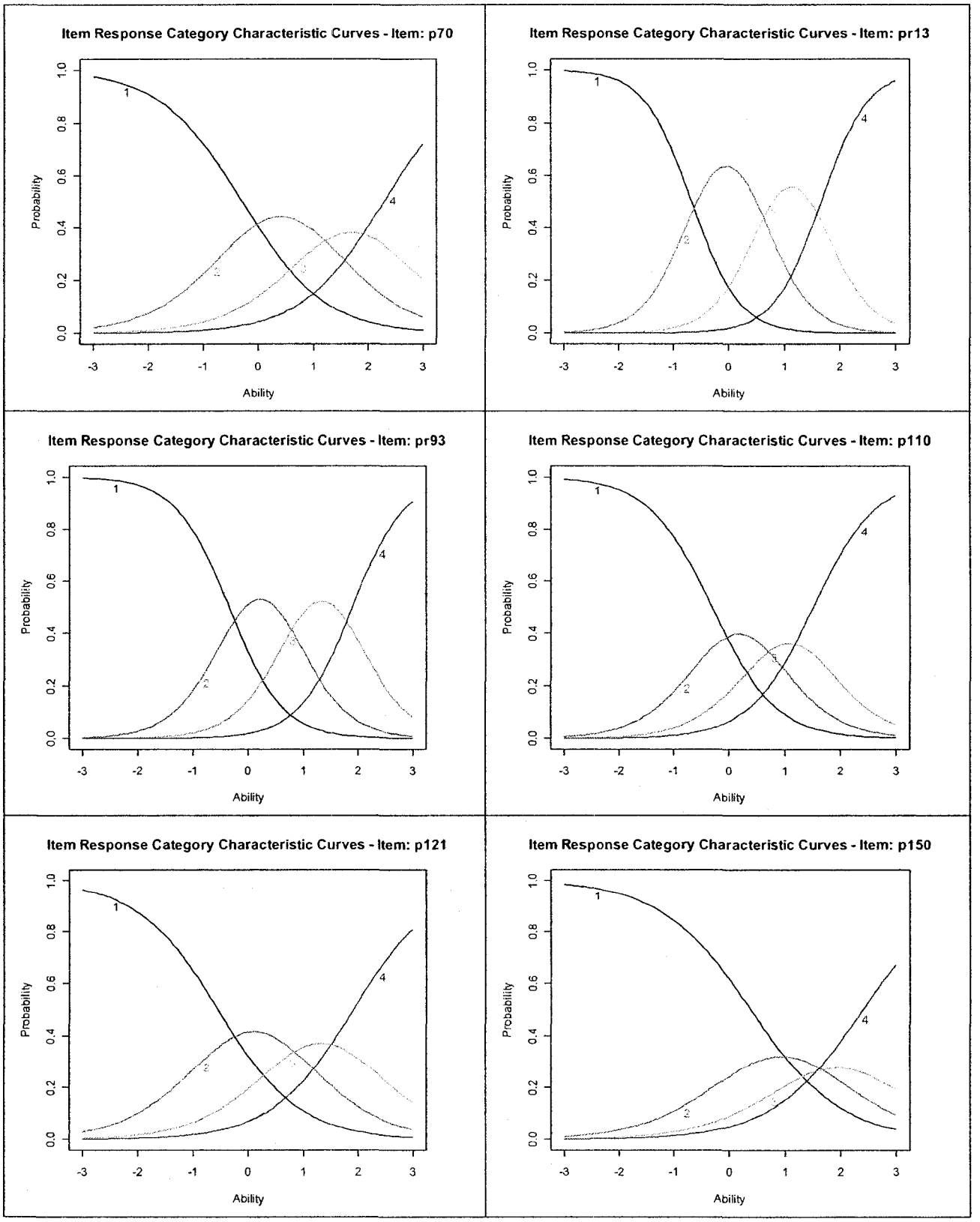

Figure 8 cont'd. Item Response Category Characteristic Curves (CCC) for SZD Items (7-12)

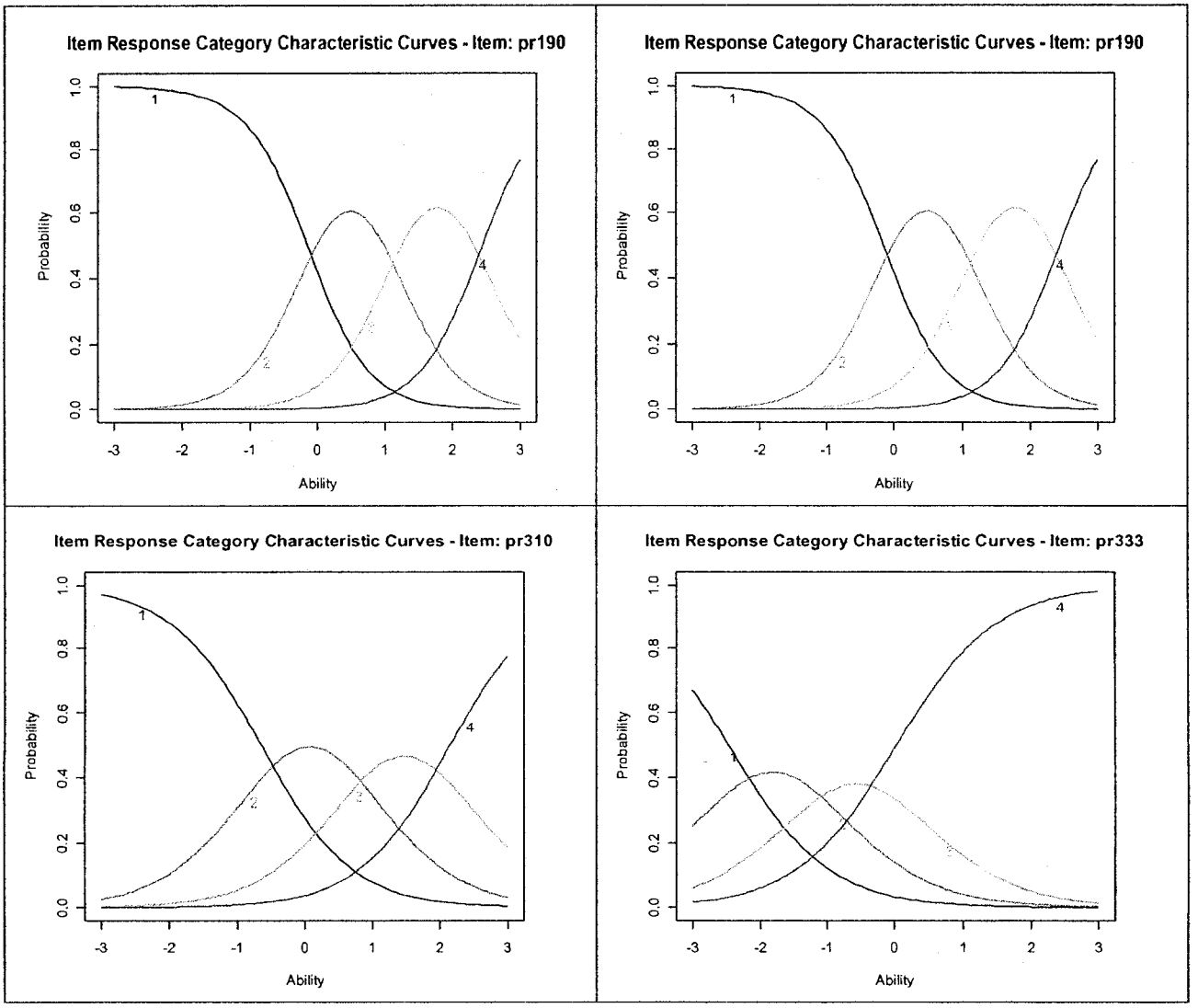

Figure 8 cont'd. Item Response Category Characteristic Curves (CCC) for SZD Items (13-16)

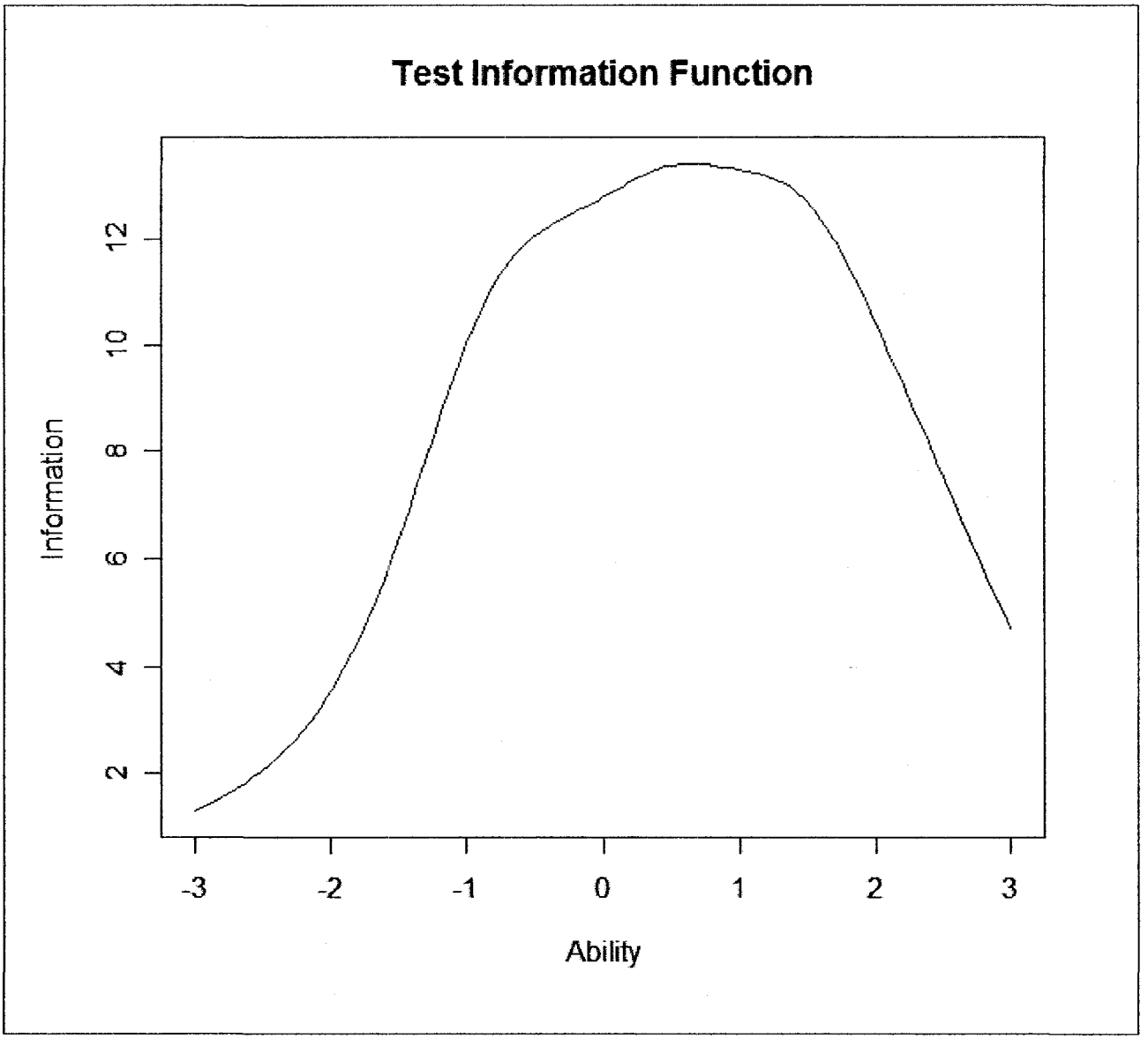

Figure 9. Test Information Function for the SZD scale.

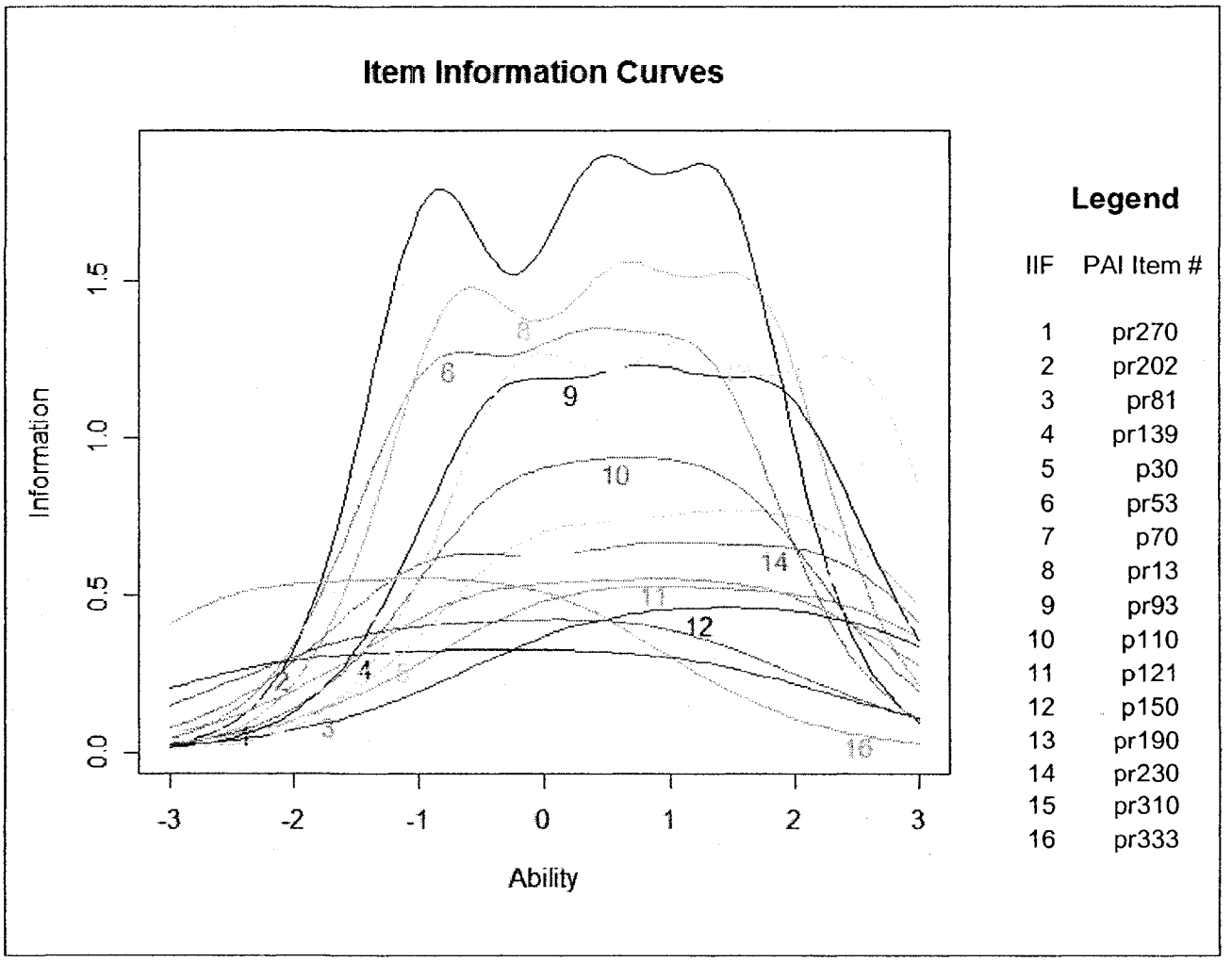

Figure 10. Item Information Functions for the SZD scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 9.1

SZD: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
pr270	-0.89	0.41	1.40	2.64	.53
pr202	-1.35	0.09	1.00	1.18	.44
pr81	0.02	1.04	2.36	1.31	.44
pr139	-1.94	-0.35	1.07	1.04	.38
p30	-0.04	1.27	2.26	1.60	.48
pr53	-0.87	0.29	1.27	2.16	.48
p70	-0.30	1.10	2.29	1.35	.44
pr13	-0.67	0.60	1.67	2.37	.51
pr93	-0.35	0.79	1.90	2.07	.47
p110	-0.31	0.65	1.52	1.75	.48
p121	-0.55	0.76	1.92	1.35	.45
p150	0.39	1.48	2.41	1.22	.45
pr190	-0.15	1.13	2.45	2.18	.50
pr230	-0.85	0.81	2.22	1.53	.42
pr310	-0.65	0.80	2.16	1.49	.47
pr333	-2.47	-1.16	0.03	1.35	.45
Mean	-0.69	.61	1.75	1.66	*(.46)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 9.2

SZD: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information (%)
-3 to +3	91.62
-2 to +2	75.04
-3 to 0	35.02
0 to +3	56.60
-3 to -2	3.70
-2 to -1	11.13
-1 to 0	20.19
0 to +1	22.59
+1 to +2	21.13
+2 to +3	12.88

Note. [1] = SZD trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 9.3

SZD: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information (%)
pr270	11.42
pr202	3.73
pr81	4.74
pr139	3.48
p30	5.72
pr53	8.52
p70	4.74
pr13	10.00
pr93	8.20
p110	5.91
p121	4.63
p150	3.71
pr190	9.29
pr230	6.03
pr310	5.61
pr333	4.67

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix F

SZT

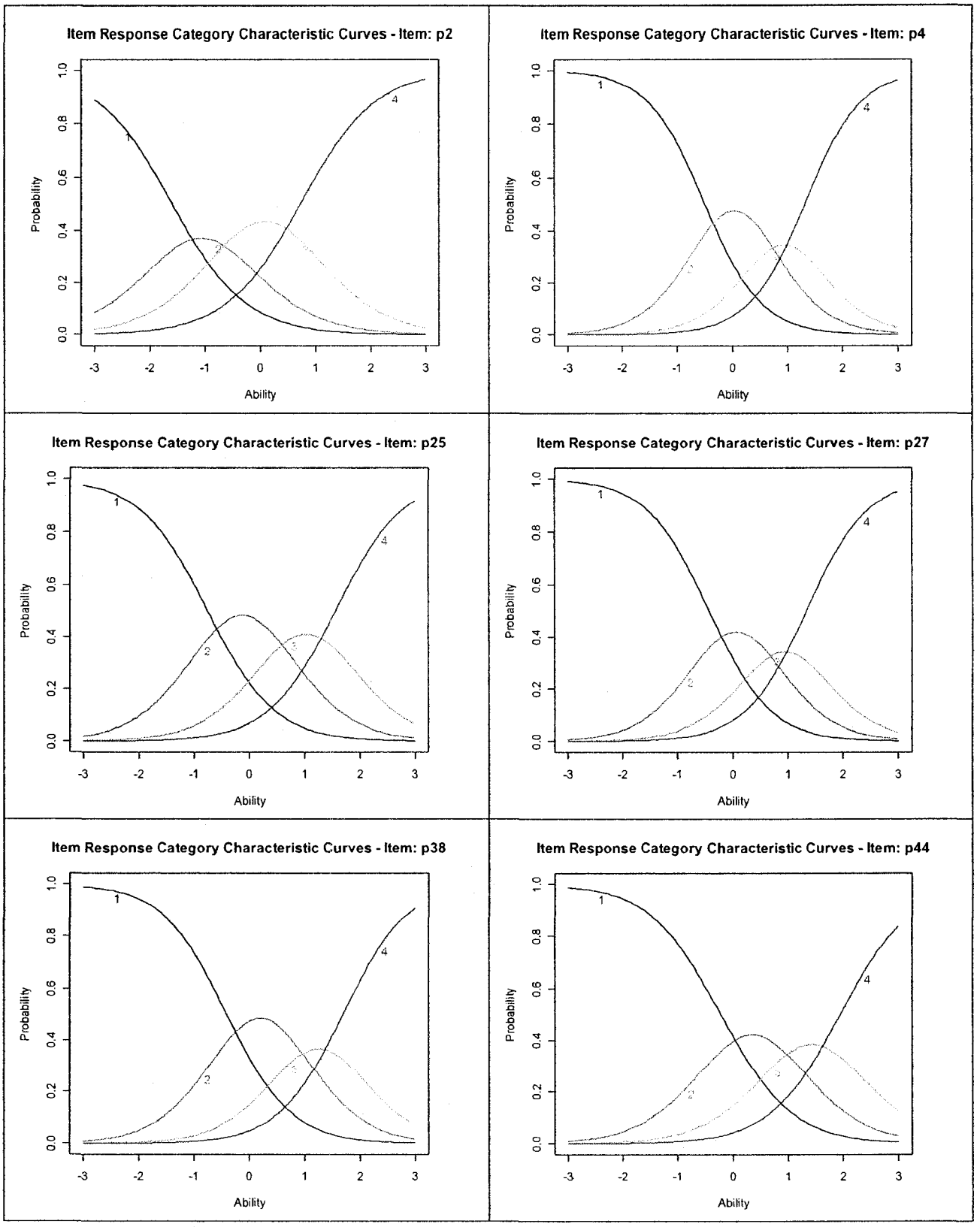

Figure 11. Item Response Category Characteristic Curves (CCC) for SZT Items (1-6)

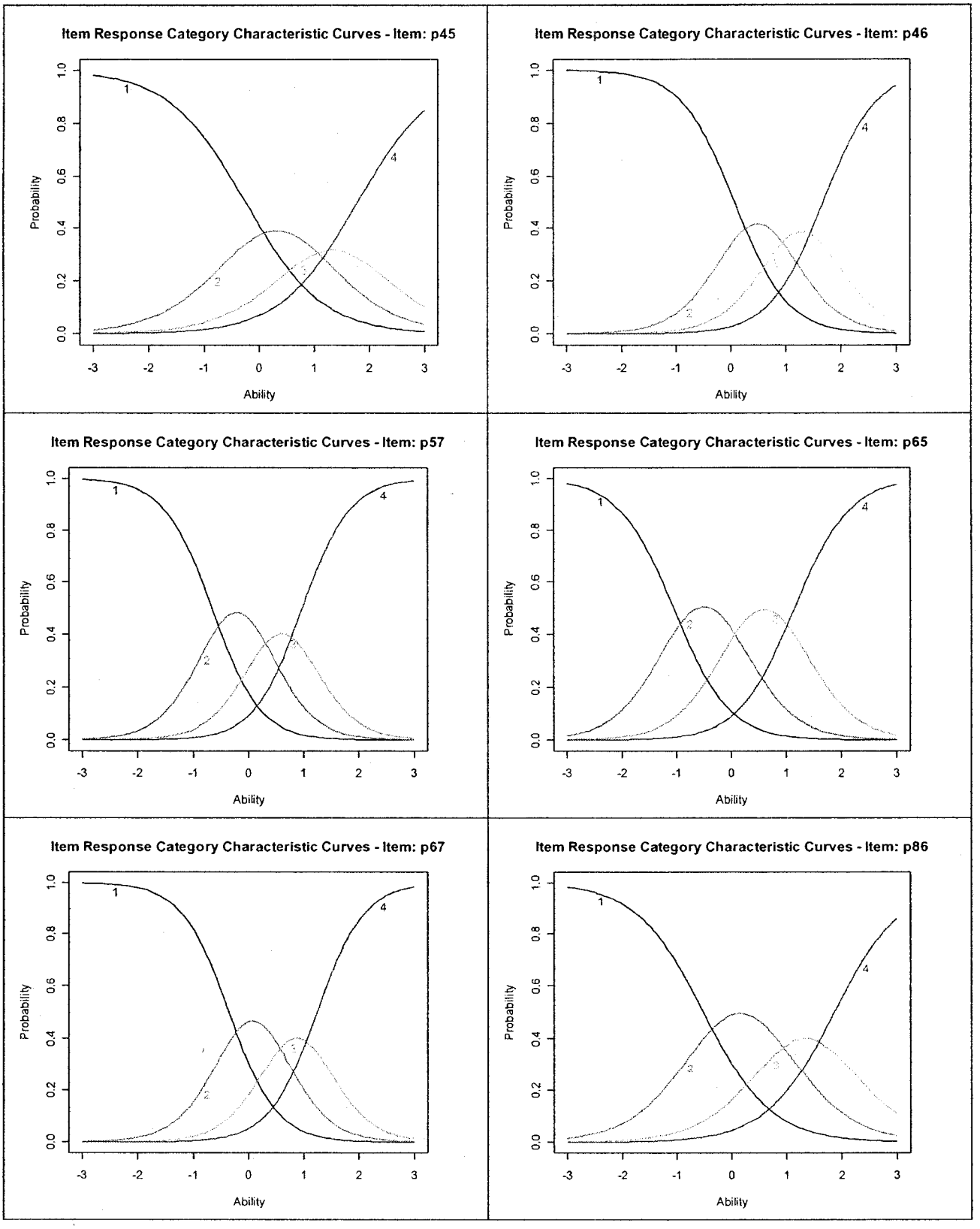

Figure 11 cont'd. Item Response Category Characteristic Curves (CCC) for SZT Items (7-12)

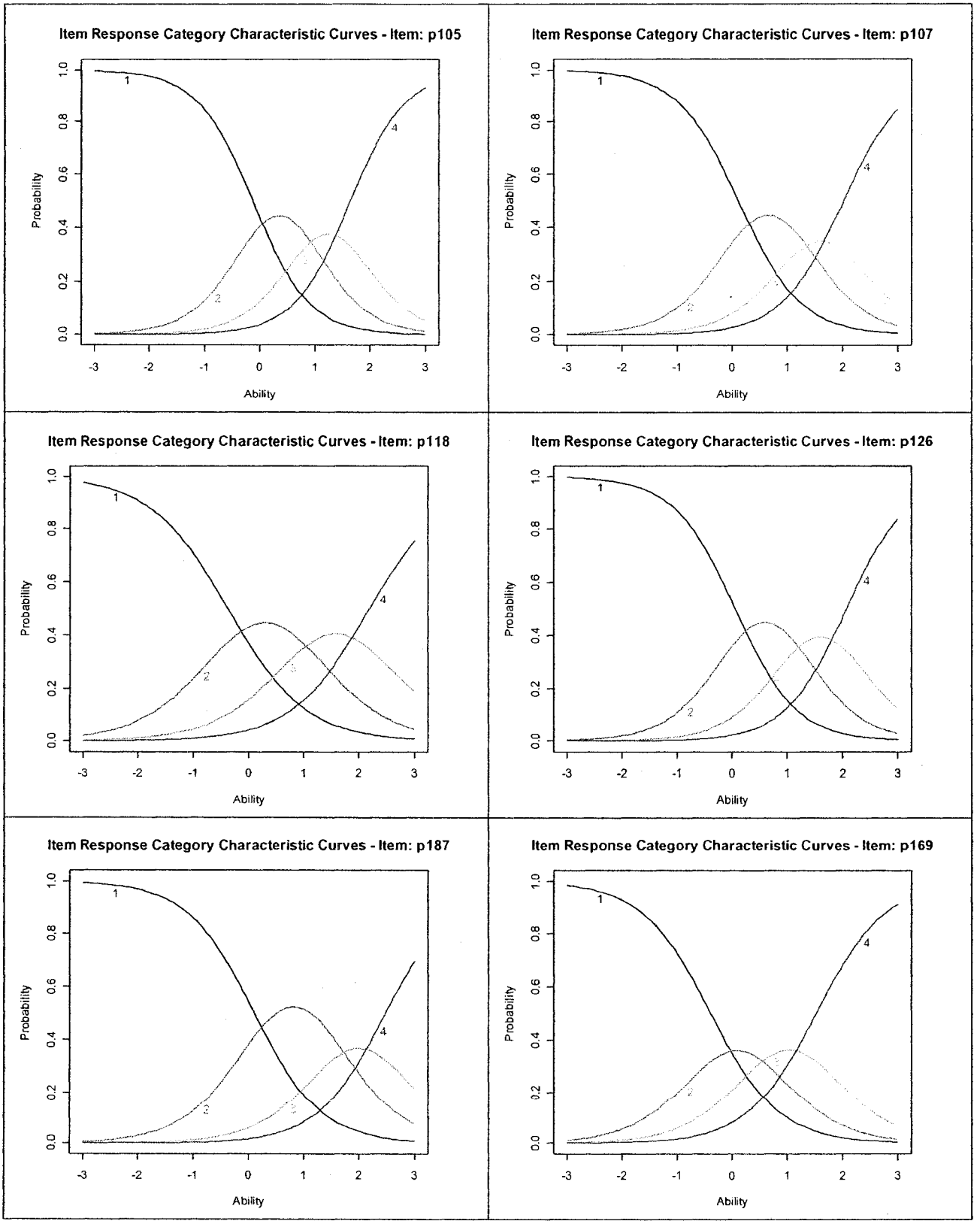

Figure 11 cont'd. Item Response Category Characteristic Curves (CCC) for SZT Items (13-18)

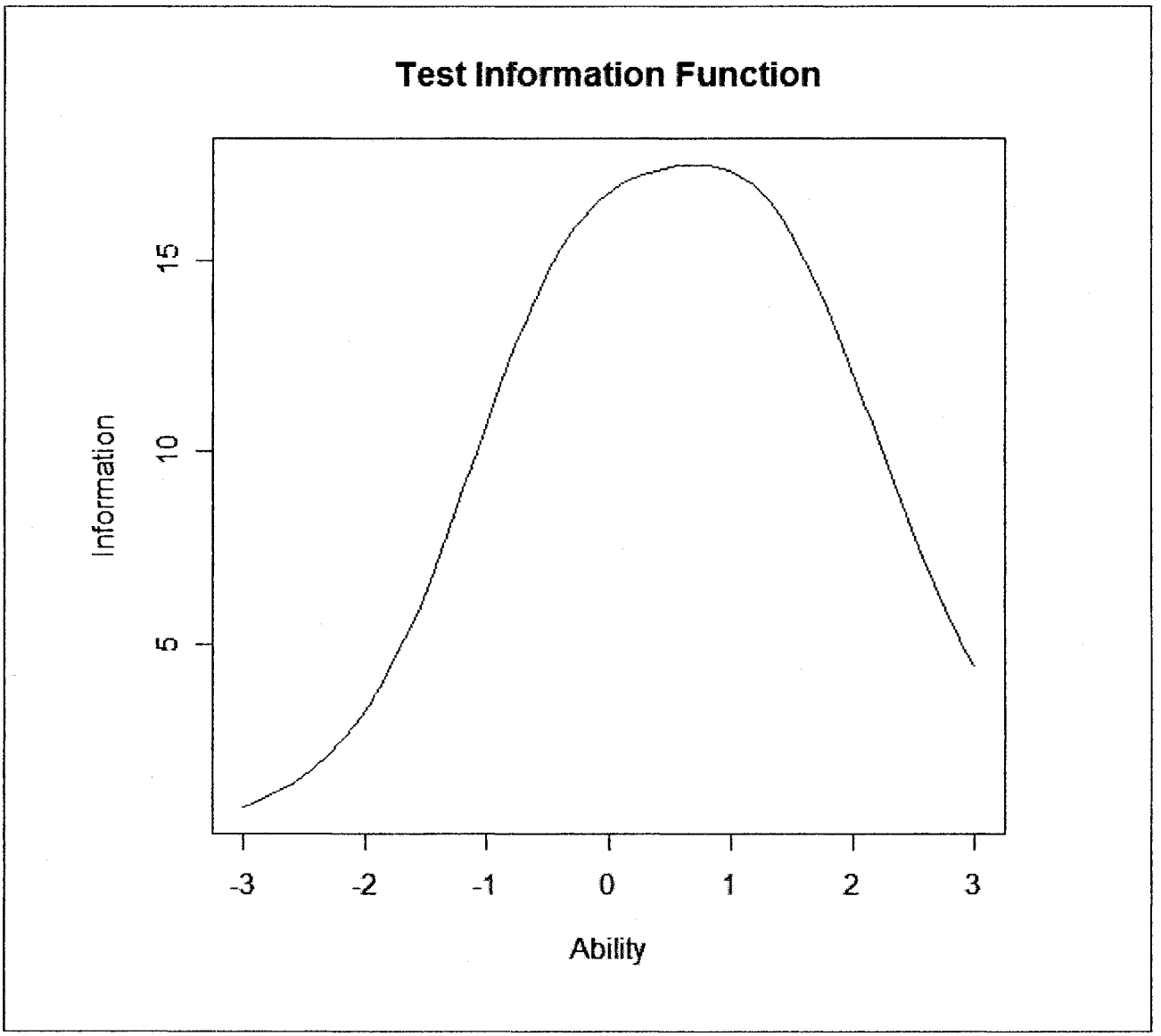

Figure 12. Test Information Function for the SZT scale.

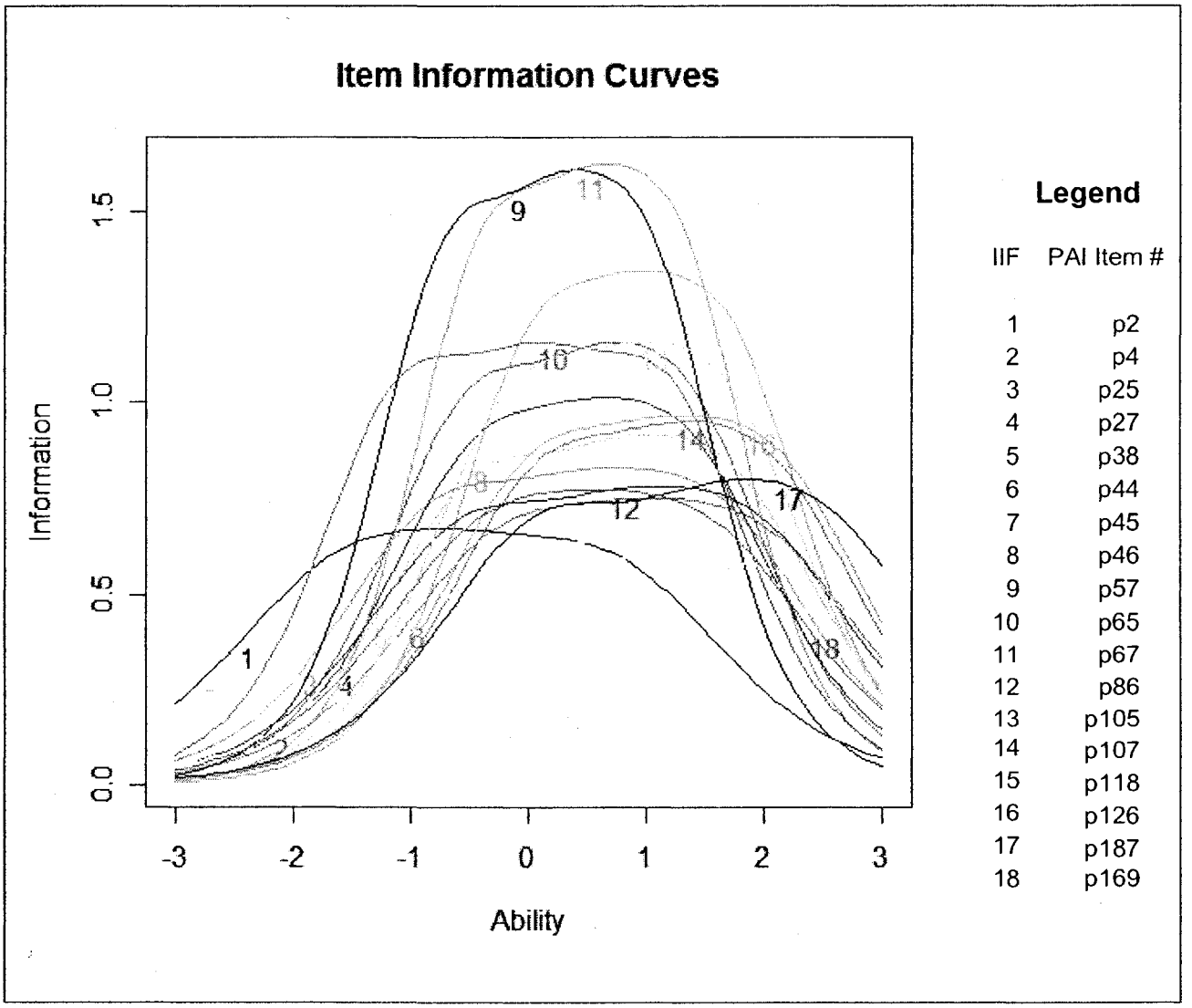

Figure 13. Item Information Functions for the SZT scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 10.1

SZT: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p2	-1.59	-0.54	0.72	1.49	.51
p4	-0.49	0.57	1.32	1.95	.53
p25	-0.75	0.51	1.56	1.67	.52
p27	-0.42	0.56	1.35	1.82	.52
p38	-0.40	0.81	1.70	1.74	.51
p44	-0.22	0.93	1.96	1.57	.51
p45	-0.25	0.89	1.79	1.45	.48
p46	0.07	0.91	1.68	2.10	.55
p57	-0.66	0.24	0.98	2.32	.57
p65	-1.05	0.07	1.16	2.00	.53
p67	-0.36	0.52	1.25	2.33	.54
p86	-0.52	0.83	1.88	1.62	.49
p105	-0.13	0.85	1.66	1.95	.53
p107	0.12	1.20	2.03	1.77	.51
p118	-0.38	0.99	2.20	1.41	.47
p126	0.06	1.15	2.08	1.79	.50
p187	0.11	1.54	2.49	1.63	.51
p169	-0.38	0.57	1.53	1.59	.49
Mean	-0.40	0.70	1.63	1.79	*(.51)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 10.2

SZT: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	94.64
-2 to +2	80.06
-3 to 0	33.88
0 to +3	60.75
-3 to -2	2.59
-2 to -1	9.83
-1 to 0	21.46
0 to +1	25.86
+1 to +2	22.91
+2 to +3	11.98

Note. [1] = SZT trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 10.3

SZT: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p2	4.55
p4	6.02
p25	5.34
p27	5.40
p38	5.45
p44	4.81
p45	4.16
p46	6.39
p57	7.38
p65	6.77
p67	7.35
p86	5.19
p105	6.00
p107	5.34
p118	4.43
p126	5.58
p187	5.22
p169	4.61

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix G

AVD

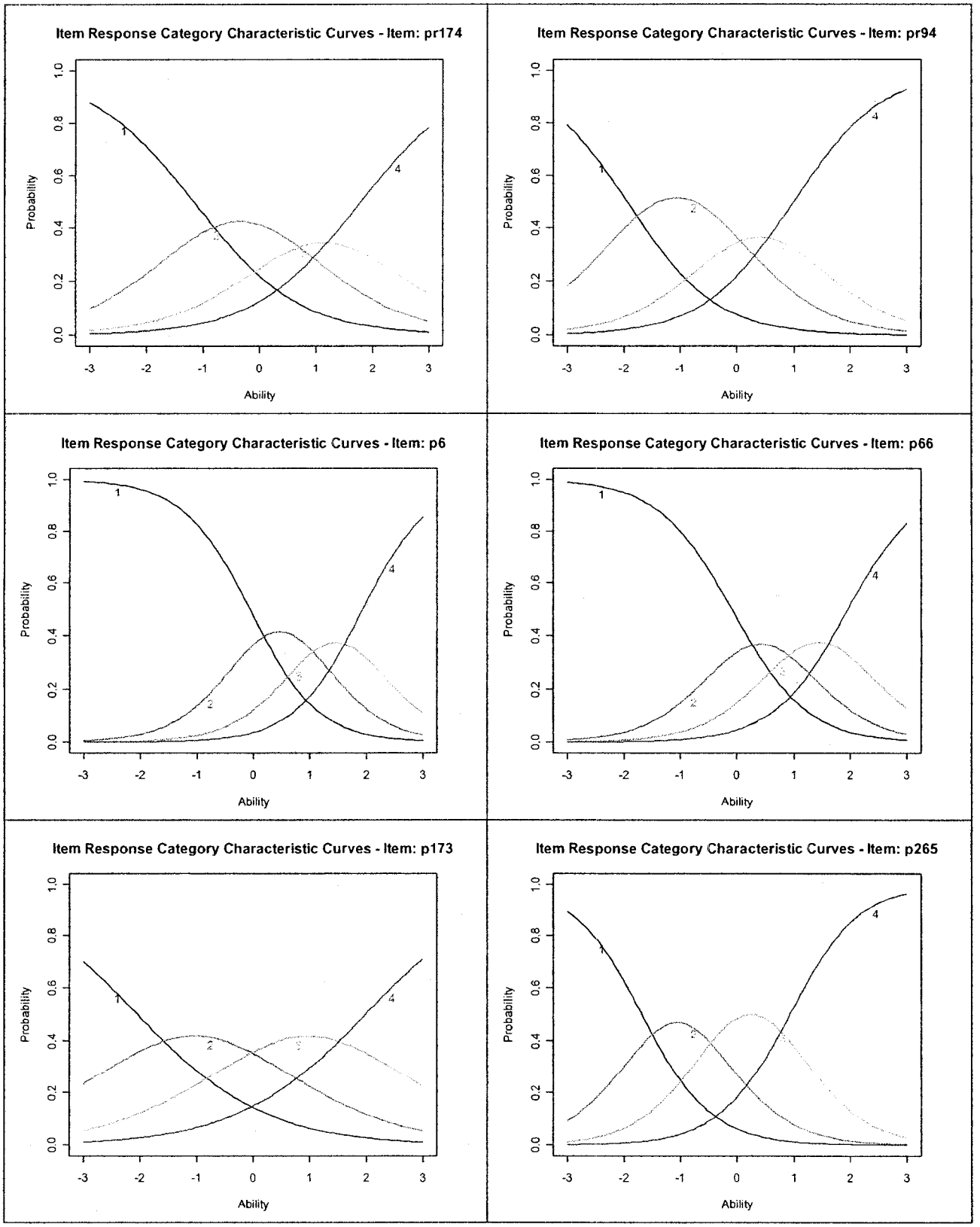

Figure 14. Item Response Category Characteristic Curves (CCC) for AVD Items (1-6)

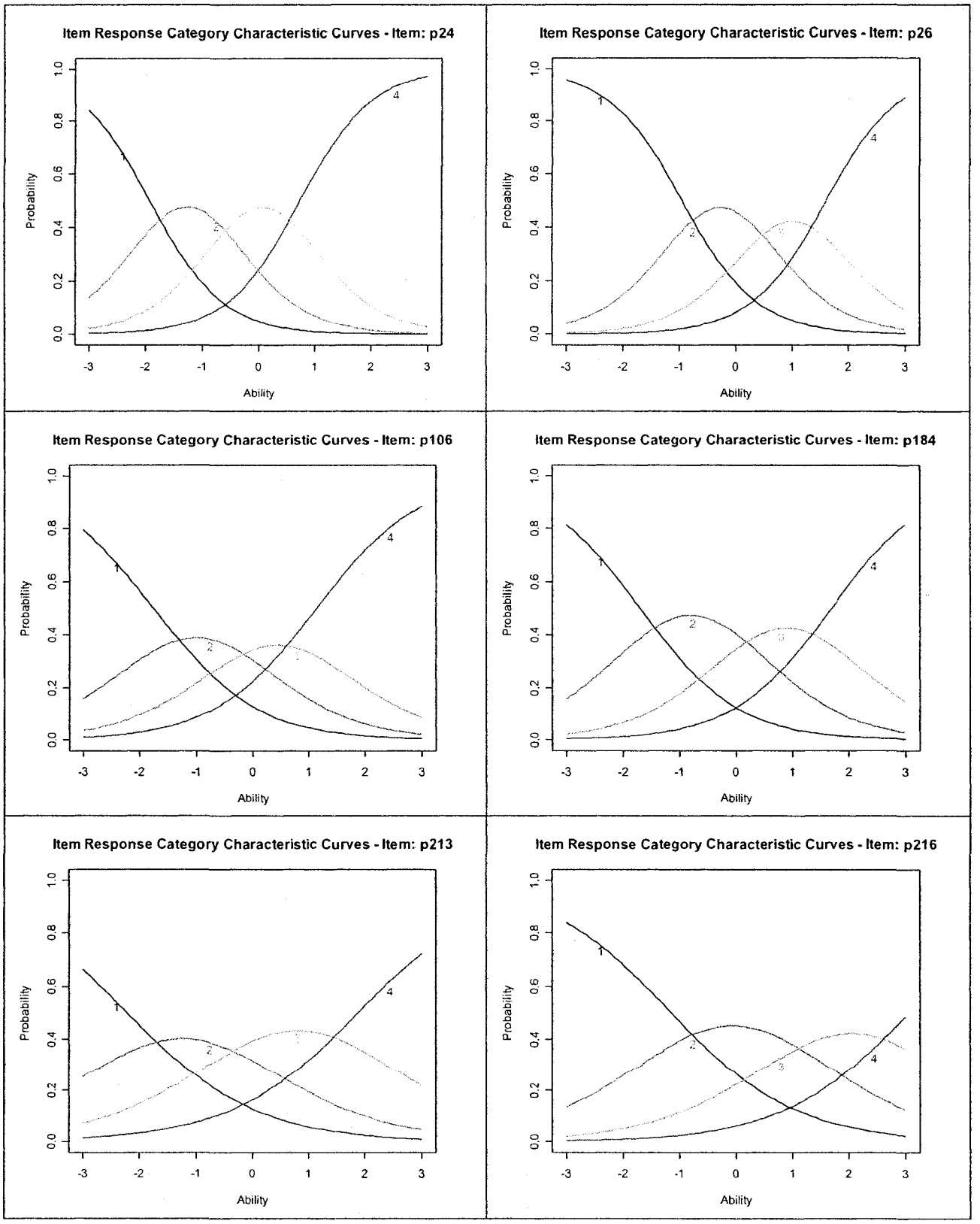

Figure 14 cont'd. Item Response Category Characteristic Curves (CCC) for AVD Items (7-12)

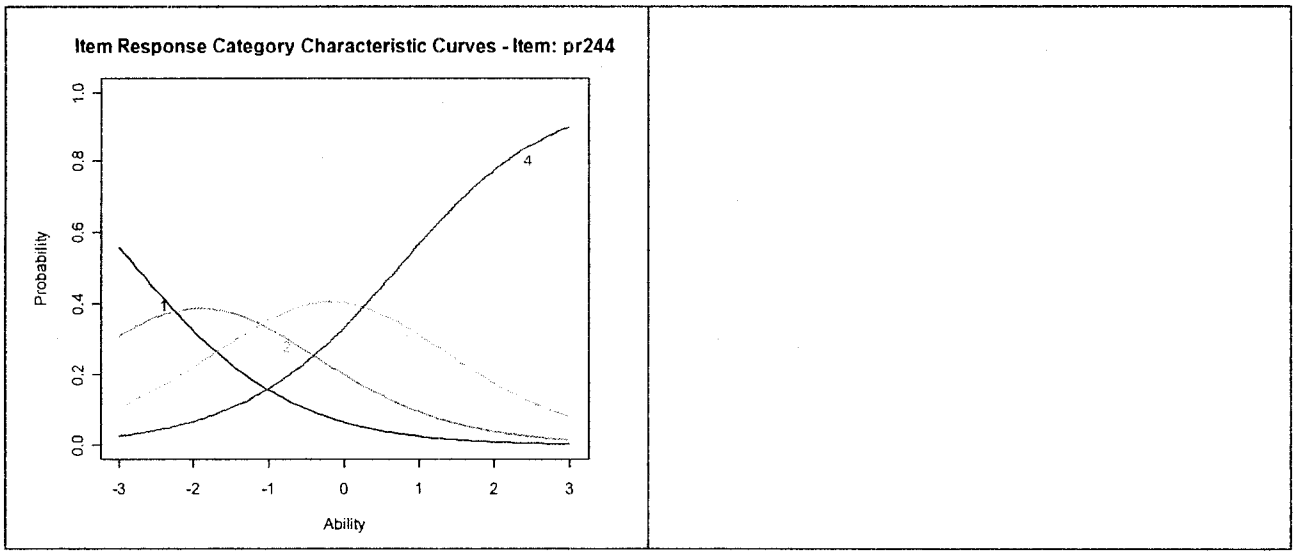

Figure 14 cont'd. Item Response Category Characteristic Curves (CCC) for AVD Items (13)

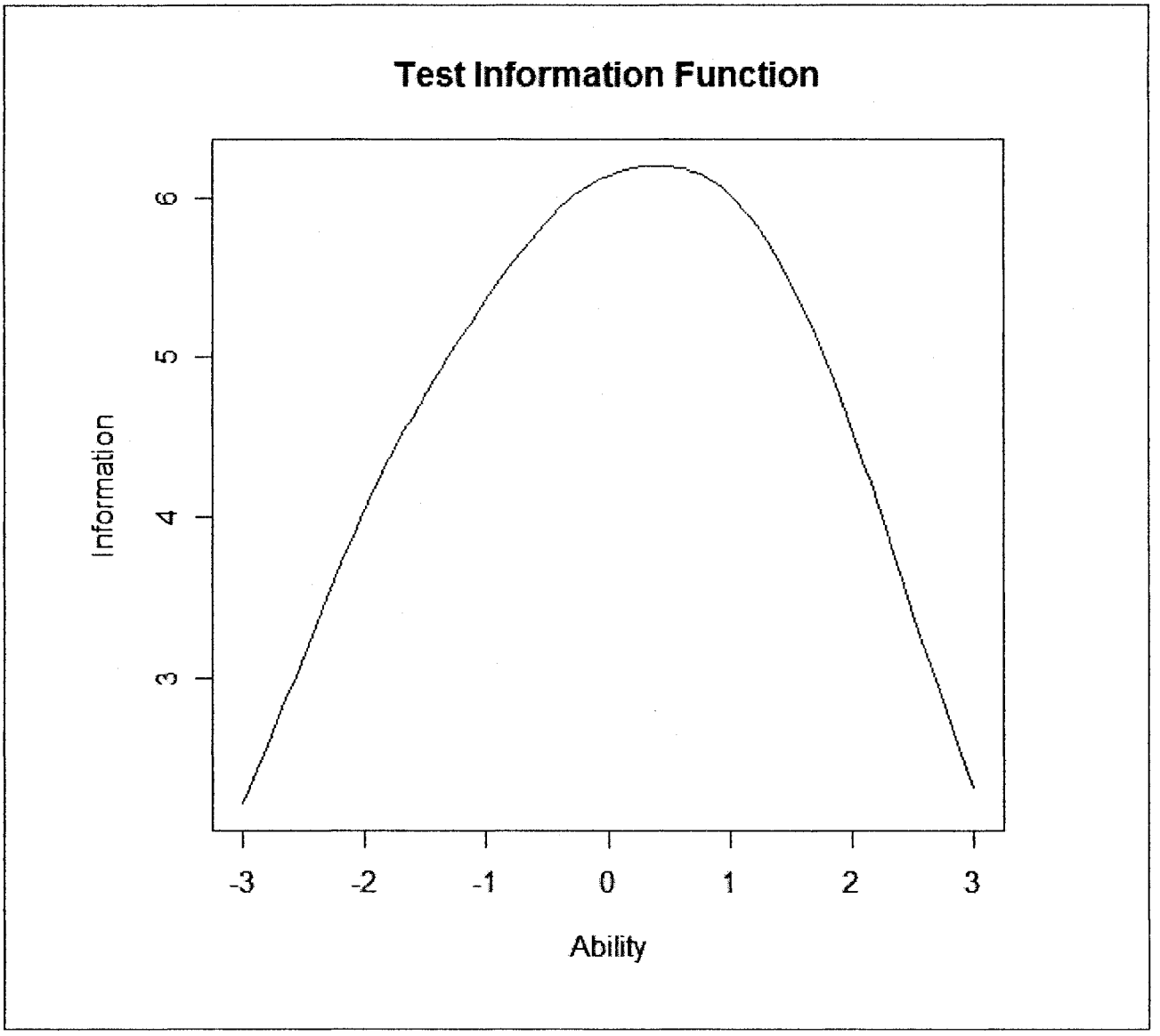

Figure 15. Test Information Function for the AVD scale.

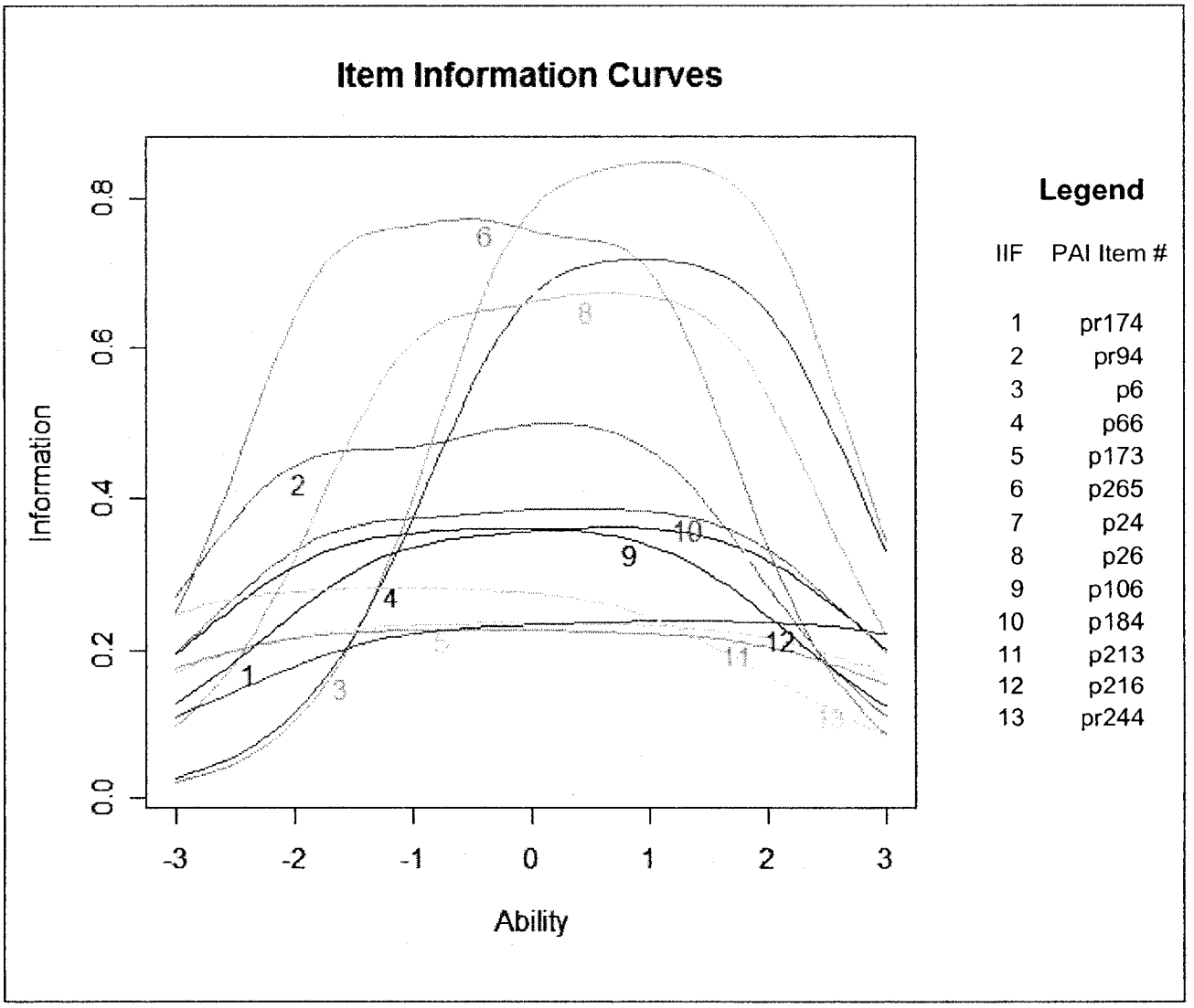

Figure 16. Item Information Functions for the AVD scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 11.1

AVD: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
pr174	-1.17	0.49	1.80	1.09	.34
pr94	-1.95	-0.18	1.00	1.29	.34
p6	-0.06	0.99	1.93	1.67	.40
p66	-0.09	0.92	1.96	1.53	.35
p173	-2.04	-0.03	1.98	0.89	.31
p265	-1.68	-0.43	0.93	1.62	.41
p24	-1.92	-0.59	0.74	1.55	.37
p26	-0.95	0.42	1.61	1.50	.39
p106	-1.75	-0.25	1.14	1.09	.32
p184	-1.70	0.11	1.70	1.14	.34
p213	-2.21	-0.26	1.88	0.87	.31
p216	-1.14	1.04	3.06	0.89	.32
pr244	-2.76	-1.06	0.72	0.97	.29
Mean	-1.49	0.09	1.57	1.24	*(.35)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 11.2

AVD: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	85.23
-2 to +2	65.79
-3 to 0	40.68
0 to +3	44.55
-3 to -2	9.29
-2 to -1	14.09
-1 to 0	17.3
0 to +1	18.33
+1 to +2	16.07
+2 to +3	10.15

Note. [1] = AVD trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 11.3

AVD: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
pr174	6.45
pr94	8.12
p6	10.02
p66	8.92
p173	5.47
p265	10.77
p24	10.20
p26	9.58
p106	6.37
p184	7.29
p213	5.35
p216	5.62
pr244	5.80

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix H

ANT

Appendix H.1

IRT results for the original ANT scales

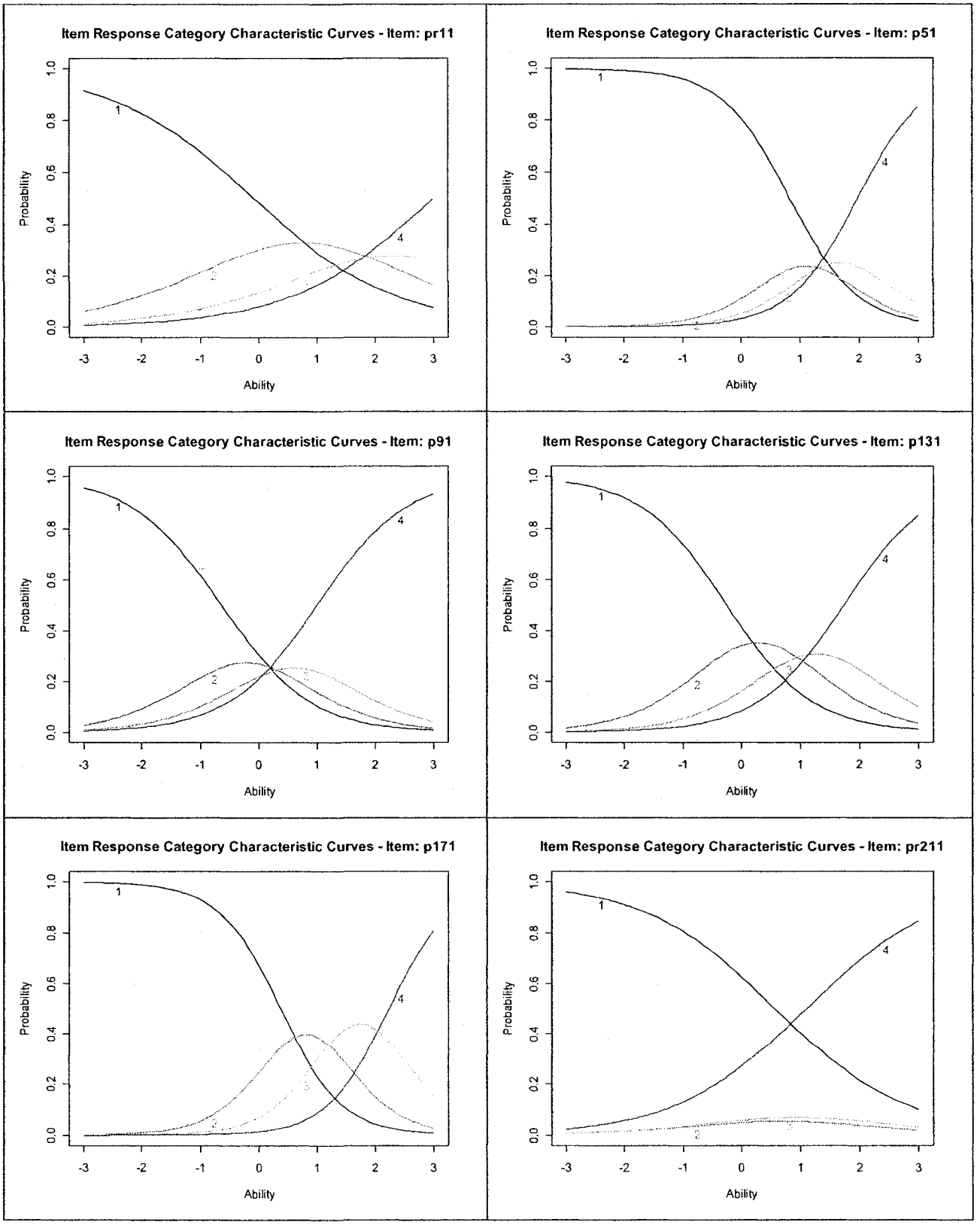

Figure 17. Item Response Category Characteristic Curves (CCC) for ANT Items (1-6)

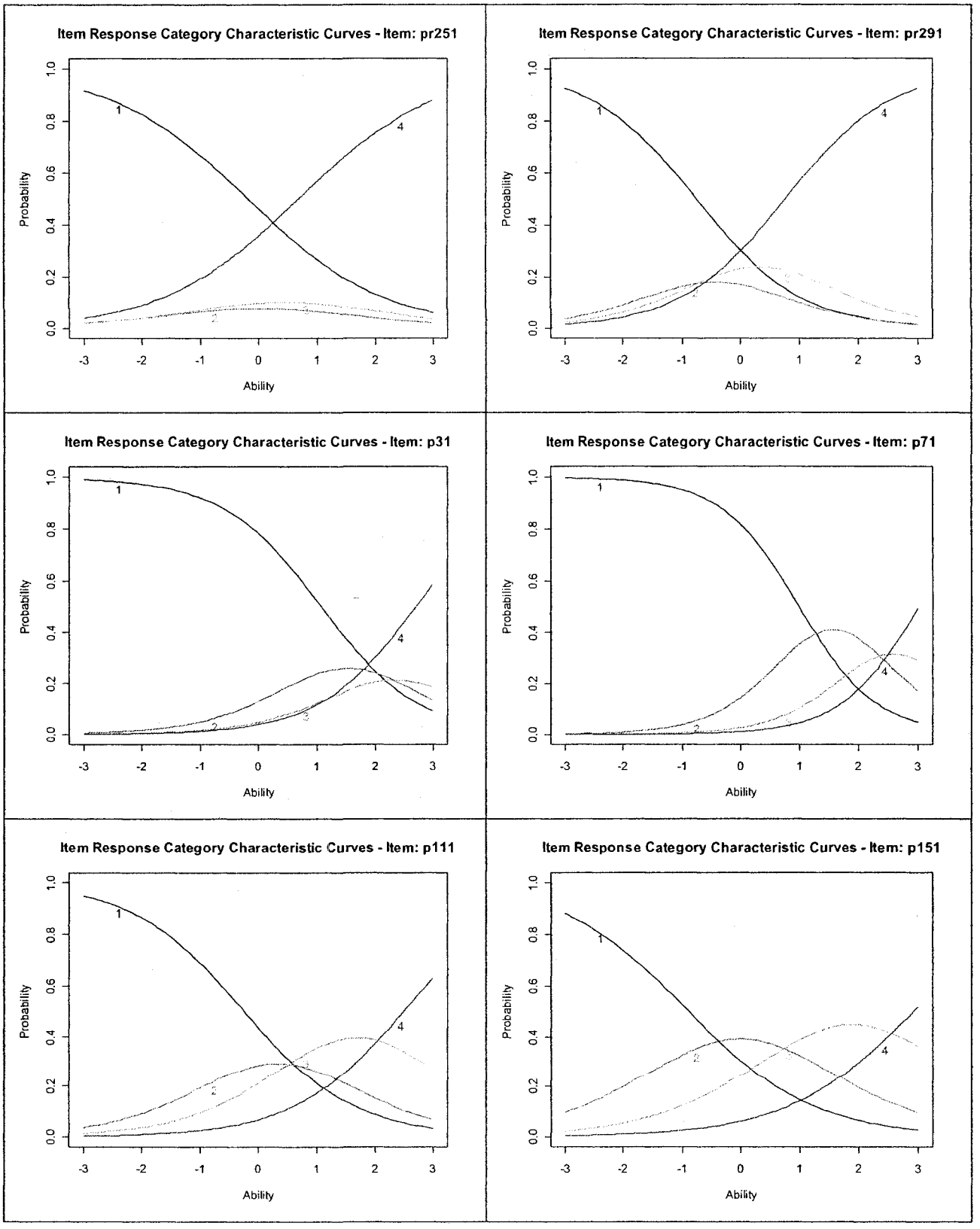

Figure 17 cont'd. Item Response Category Characteristic Curves (CCC) for ANT Items (7-12)

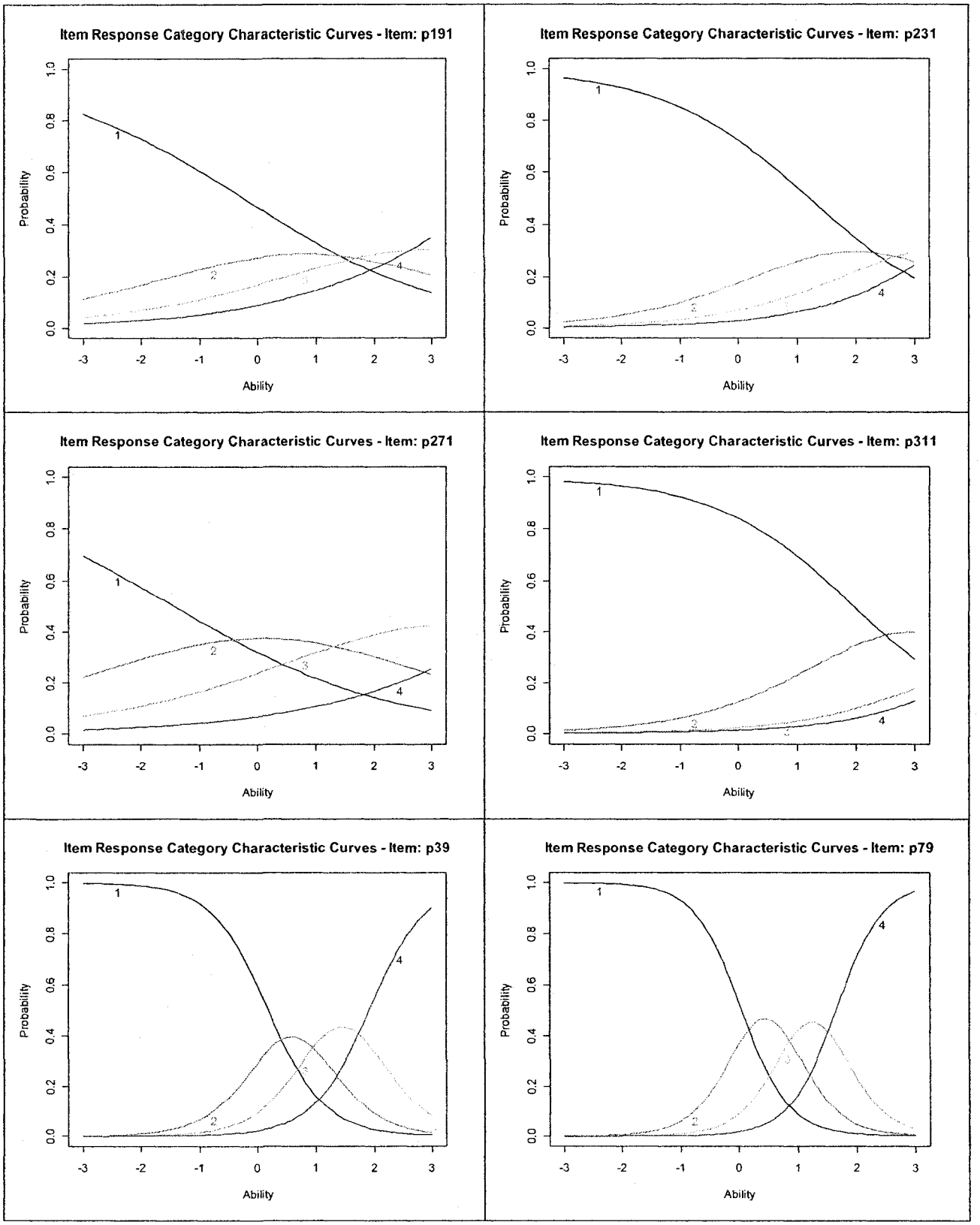

Figure 17 cont'd. Item Response Category Characteristic Curves (CCC) for ANT Items (13-18)

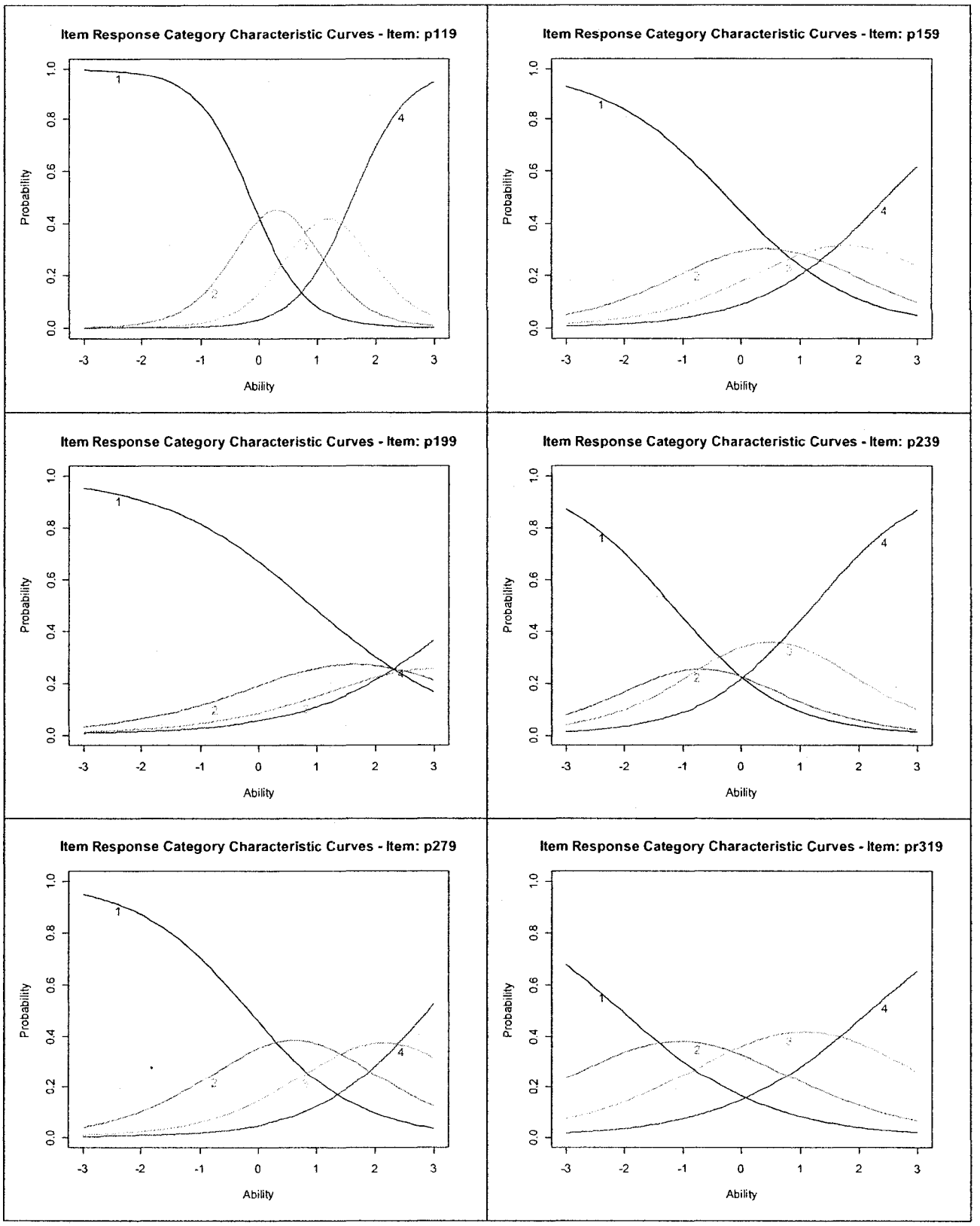

Figure 17 cont'd. Item Response Category Characteristic Curves (CCC) for ANT Items (19-24)

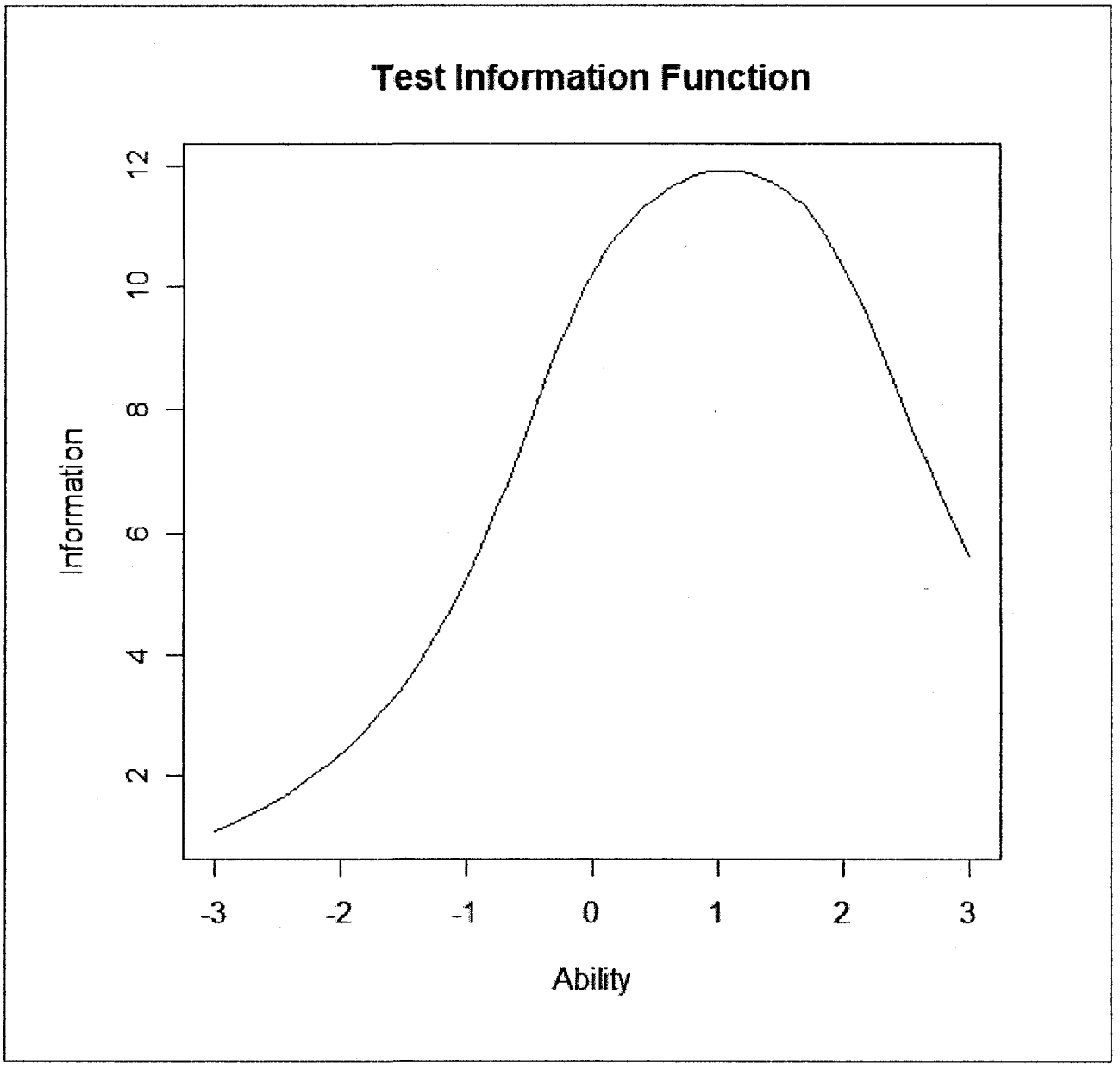

Figure 18. Test Information Function for the original ANT scale.

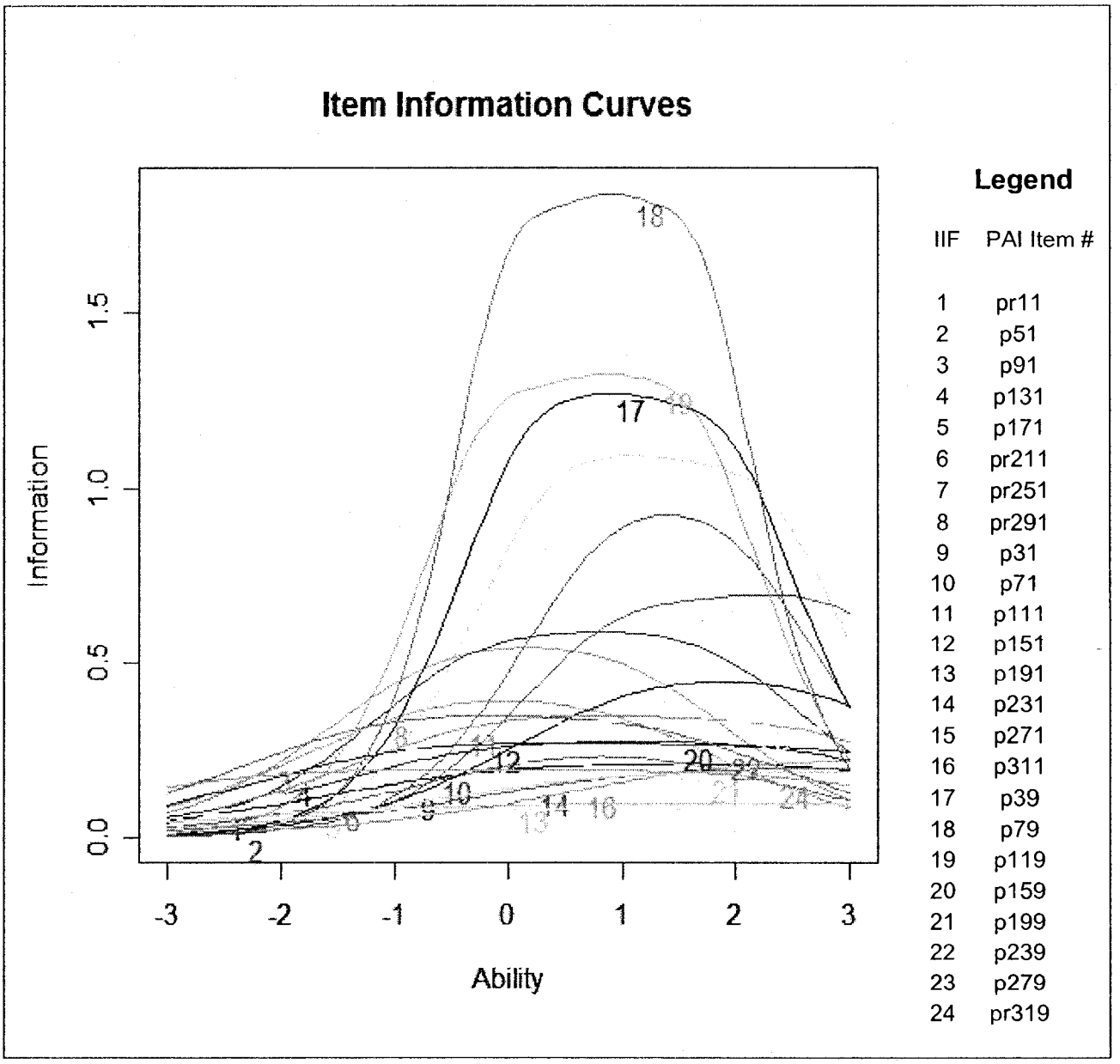

Figure 19. Item Information Functions for the original ANT scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 12.1

Original ANT: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g (H)
pr11	-0.08	1.59	3.00	0.82	.26
p51	0.83	1.39	1.98	1.73	.38
p91	-0.64	0.21	1.00	1.32	.35
p131	-0.25	0.81	1.73	1.38	.34
p171	0.37	1.25	2.24	1.91	.38
pr211	0.56	0.80	1.10	0.91	.28
pr251	-0.18	0.19	0.68	0.86	.28
pr291	-0.75	-0.10	0.77	1.12	.32
p31	1.08	1.97	2.70	1.19	.34
p71	0.99	2.15	3.02	1.51	.38
p111	-0.25	0.90	2.49	1.06	.27
p151	-0.89	0.87	2.92	0.94	.29
p191	-0.25	1.86	4.11	0.56	.22
p231	1.21	2.77	4.41	0.79	.28
p271	-1.43	1.58	5.06	0.53	.23
p311	1.96	3.96	5.25	0.85	.32
p39	0.18	1.00	1.91	2.05	.37
p79	0.04	0.86	1.64	2.50	.39
p119	-0.15	0.78	1.62	2.10	.38
p159	-0.24	1.10	2.49	0.94	.29
p199	0.91	2.35	3.71	0.77	.28
p239	-1.19	-0.20	1.22	1.05	.27
p279	-0.17	1.39	2.90	1.04	.25
pr319	-2.06	-0.05	2.20	0.80	.25
Mean	-0.02	1.23	2.51	1.20	*(.31)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 12.2

Original ANT: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	81.79
-2 to +2	63.90
-3 to 0	24.26
0 to +3	57.53
-3 to -2	3.04
-2 to -1	6.70
-1 to 0	14.52
0 to +1	21.21
+1 to +2	21.46
+2 to +3	14.85

Note. [1] = ANT trait range in *SD* units, *M* = 0, *SD* = 1; % = percent of total information or total area under the Test Information Function.

Table 12.3

Original ANT: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
pr11	2.75
p51	5.31
p91	4.17
p131	4.79
p171	7.39
pr211	1.98
pr251	1.98
pr291	3.26
p31	3.63
p71	5.50
p111	3.76
p151	3.67
p191	1.83
p231	2.66
p271	1.93
p311	2.94
p39	7.95
p79	10.18
p119	8.31
p159	3.16
p199	2.43
p239	3.55
p279	3.87
pr319	3.01

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix H.2

Original ANT Subscales

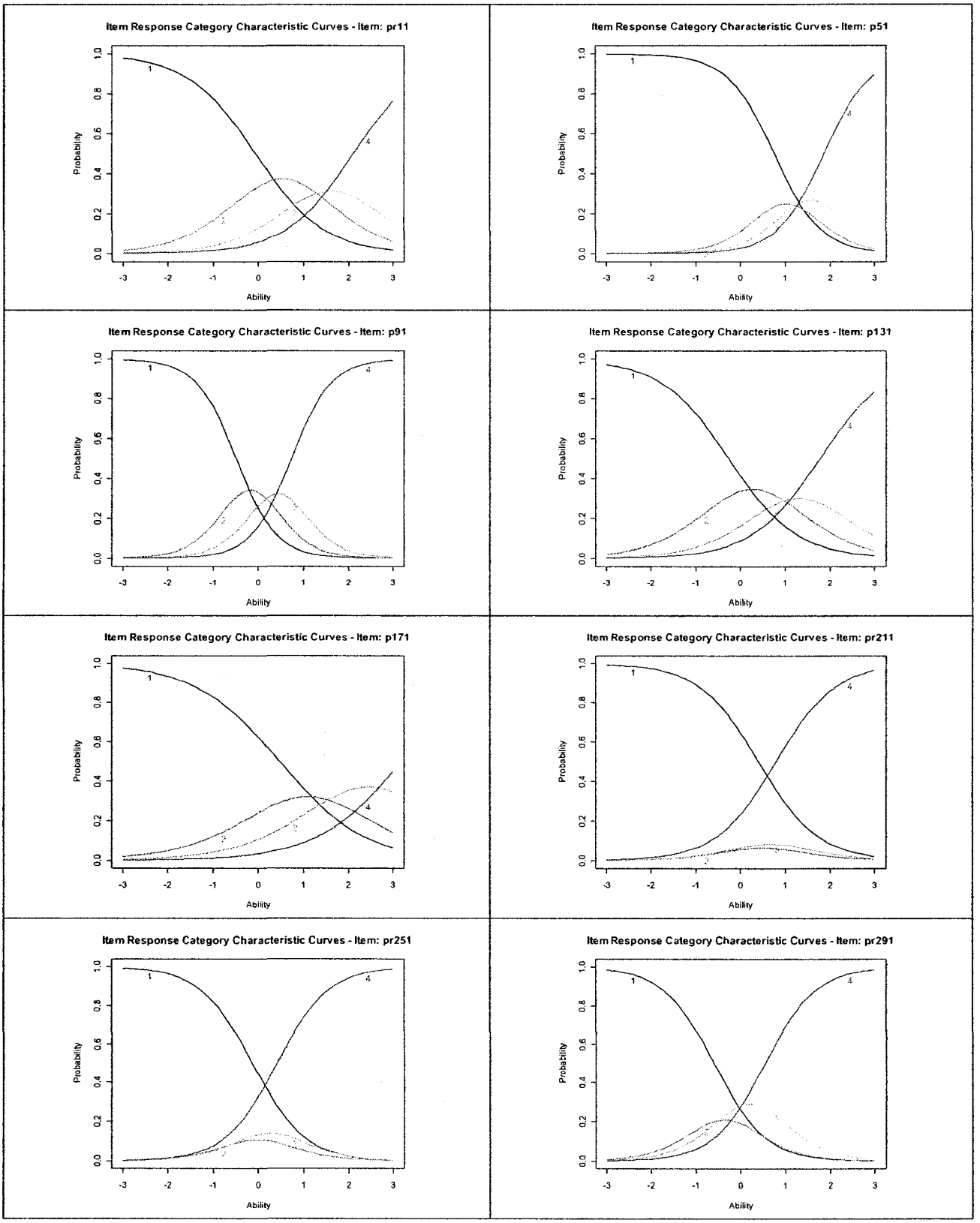

Figure 20. Item Response Category Characteristic Curves (CCC) for ANT-A Items (1-8)

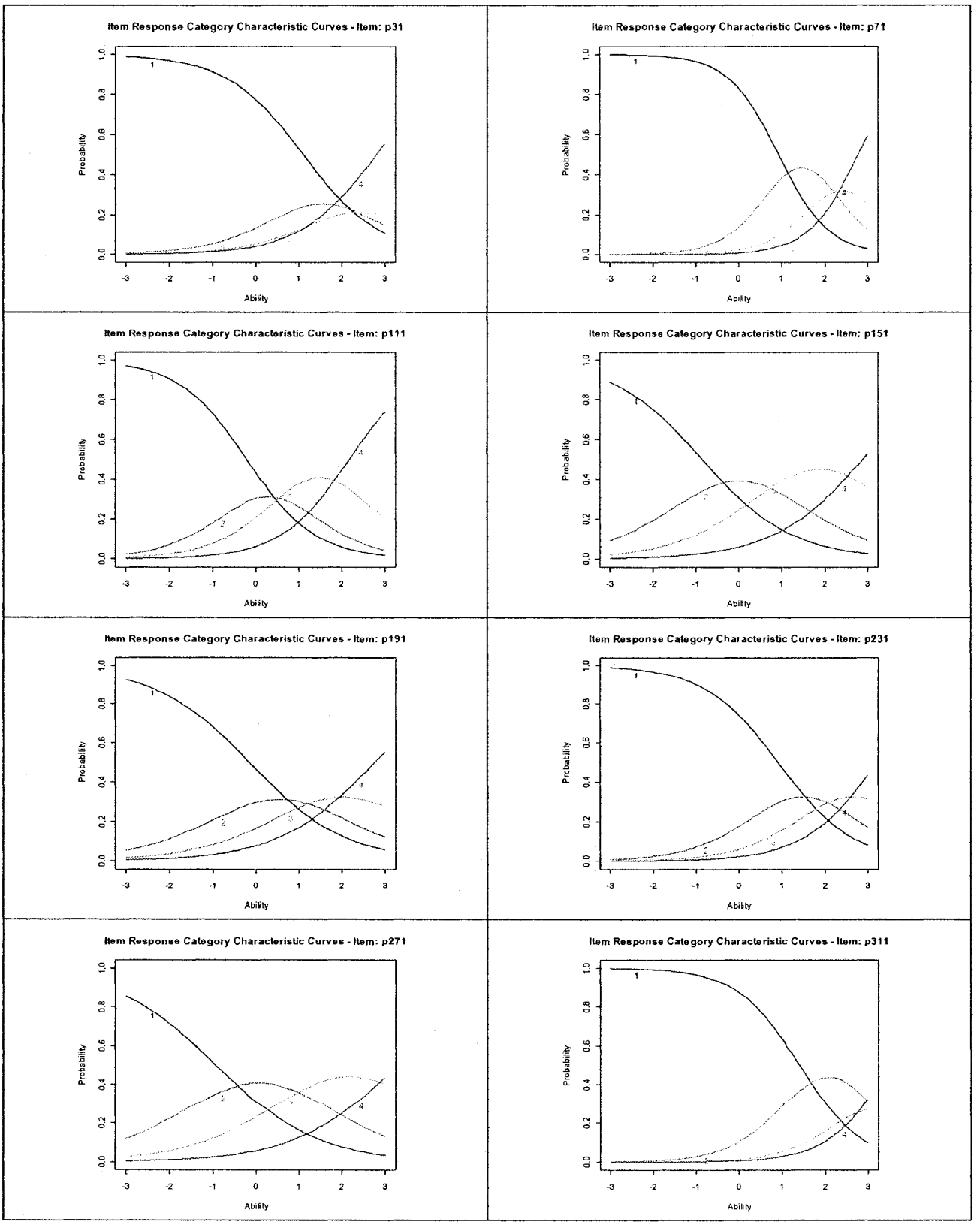

Figure 21. Item Response Category Characteristic Curves (CCC) for ANT-E Items (1-8)

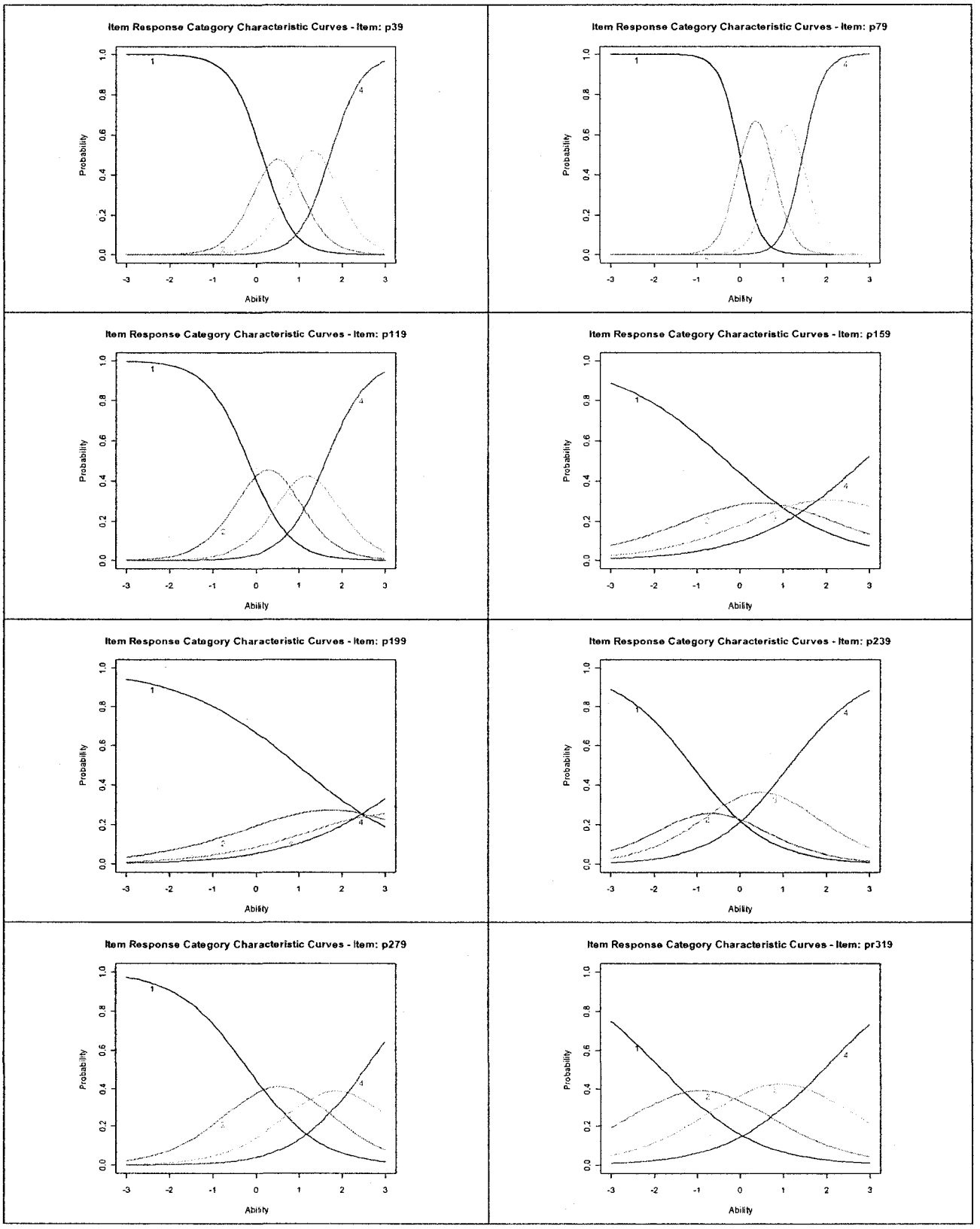

Figure 22. Item Response Category Characteristic Curves (CCC) for ANT-S Items (1-8)

Table 12.4

Legend for Figure 23: Corresponding PAI items for the IIF's of the original ANT Subscales.

Subscale IIF#[a]	Total Scale IIF#	PAI Items and Original Item Numbers
ANT-A		
3	(3)	91. I've done some things that weren't exactly legal.
2	(2)	51. I've deliberately damaged someone's property.
8	(8)	291. I've never taken money or property that wasn't mine.
1	(1)	11. I was usually well-behaved at school.
4	(4)	131. I used to lie a lot to get out of tight situations.
7	(7)	251. I've never been in trouble with the law.
5	(5)	171. I like to see how much I can get away with.
6	(6)	211. I was never expelled or suspended from school when I was young.
ANT-E		
2	(10)	71. I'll take advantage of others if they leave themselves open to it.
8	(16)	311. When I make a promise, I really don't need to keep it.
3	(11)	111. I'll do most things if the price is right.
6	(14)	231. I don't like to stay in a relationship very long.
4	(12)	151. I can talk my way out of just about anything.
1	(9)	31. I've borrowed money knowing I wouldn't pay it back.
7	(15)	271. I look after myself first; let others take care of themselves.
5	(13)	191. I don't like being tied to one person.
ANT-S		
2	(18)	79. I do a lot of wild things just for the thrill of it.
1	(17)	39. I get a kick out of doing dangerous things.
3	(19)	119. My Behaviour is pretty wild at times.
7	(23)	279. I'm not a person who turns down a dare.
6	(22)	239. I like to drive fast.
8	(24)	319. I never take risks if I can avoid it.
4	(20)	159. If I get tired of a place, I just pick up and leave.
5	(21)	199. The idea of "settling down" has never appealed to me.

Note. [a]IIF# = Item Information Function number from Figure X. Items are arranged in descending order based on amount of information contributed to the respective subscale.

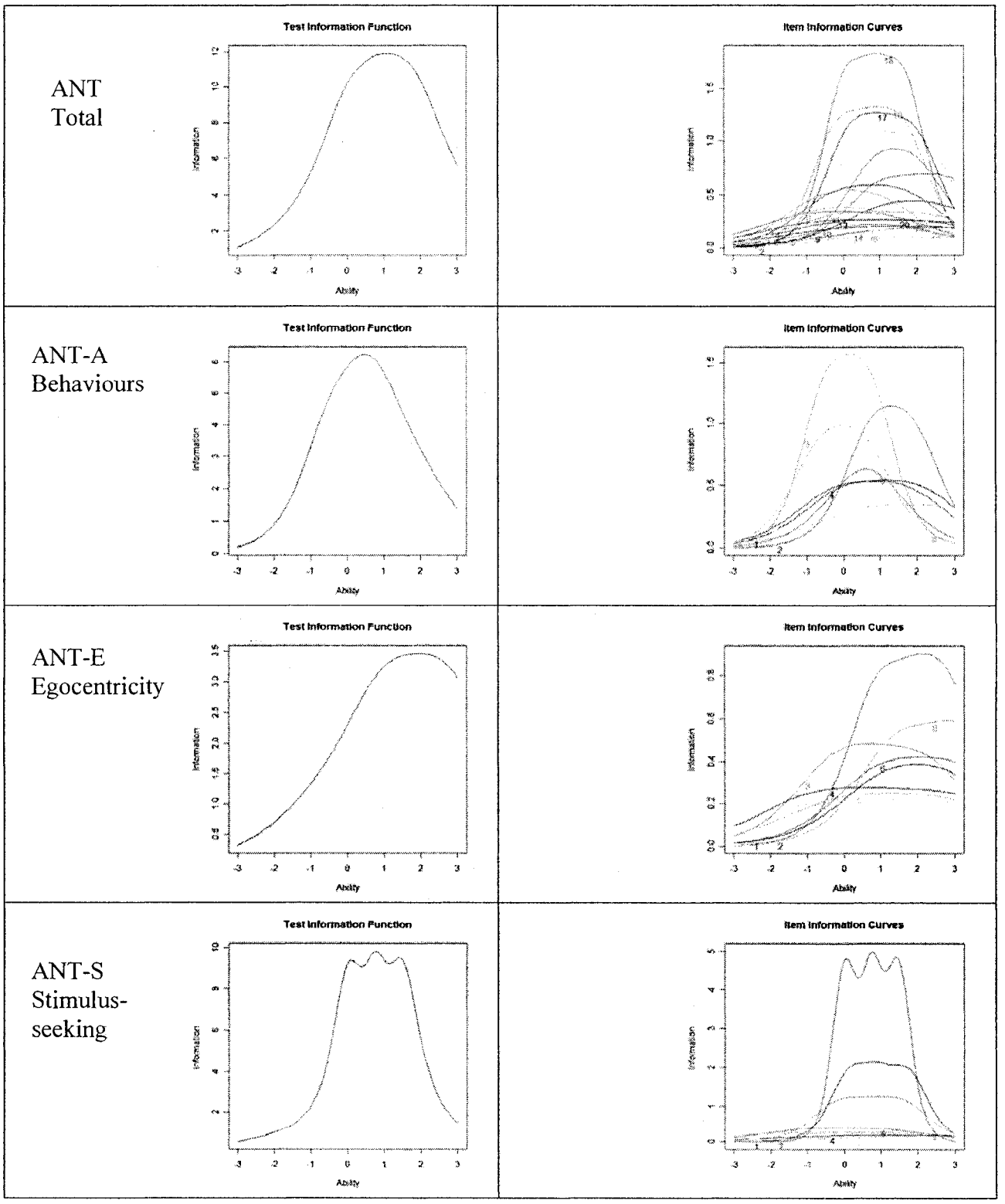

Figure 23. Test and item information functions for the original PAI ANT subscales (see Table 12.4 for a legend of the PAI items that correspond with the IIF numbers).

Table 12.5

Original ANT Subscales: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g (H)
ANT-A					
pr11	-0.06	1.13	2.12	1.32	.39
p51	0.76	1.30	1.87	1.91	.47
p91	-0.49	0.14	0.73	2.25	.46
p131	-0.27	0.83	1.78	1.31	.39
p171	0.47	1.72	3.19	1.07	.36
pr211	0.40	0.57	0.79	1.50	.39
pr251	-0.13	0.10	0.41	1.77	.42
pr291	-0.60	-0.12	0.55	1.78	.44
Mean	*0.01*	*0.71*	*1.43*	*1.61*	*(.42)*
ANT-E					
p31	1.10	2.03	2.81	1.11	.30
p71	0.93	2.01	2.78	1.72	.37
p111	-0.22	0.81	2.18	1.26	.30
p151	-0.86	0.87	2.89	0.96	.30
p191	-0.16	1.28	2.77	0.90	.29
p231	0.91	2.07	3.23	1.17	.32
p271	-0.94	1.08	3.30	0.85	.29
p311	1.40	2.74	3.54	1.39	.37
Mean	*0.27*	*1.61*	*2.94*	*1.17*	*(.31)*
ANT-S					
p39	0.13	0.89	1.74	2.72	.45
p79	0.01	0.76	1.47	4.31	.47
p119	-0.19	0.76	1.63	2.06	.41
p159	-0.32	1.23	2.88	0.77	.30
p199	0.96	2.52	3.98	0.72	.30
p239	-1.13	-0.19	1.16	1.12	.32
p279	-0.18	1.22	2.53	1.25	.32
pr319	-1.82	-0.05	1.92	0.93	.30
Mean	*-0.32*	*0.89*	*2.16*	*1.73*	*(.36)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 12.6

Original ANT Subscales: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information			
	Total Scale	ANT-A	ANT-E	ANT-S
-3 to +3	81.79	92.57	71.90	92.04
-2 to +2	63.90	79.94	50.48	80.46
-3 to 0	24.26	33.27	18.28	24.07
0 to +3	57.53	59.30	53.62	67.97
-3 to -2	3.04	2.19	2.71	2.08
-2 to -1	6.70	8.96	5.53	4.62
-1 to 0	14.52	22.13	10.03	17.37
0 to +1	21.21	28.27	15.85	30.72
+1 to +2	21.46	20.58	19.06	27.75
+2 to +3	14.85	10.44	18.70	9.50

Note. [1] = ANT trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 12.7

Original ANT Subscales: Item Information as a Function of Trait Level (Theta)

PAI Items	Percent of Total Information	
	ANT Subscale	ANT Total Scale
ANT-A		
pr11	11.72	2.75
p51	15.03	5.31
p91	19.56	4.17
p131	11.30	4.79
p171	9.43	7.39
pr211	8.36	1.98
pr251	10.88	1.98
pr291	13.73	3.26
ANT-E		
p31	10.09	3.63
p71	19.00	5.50
p111	13.62	3.76
p151	11.15	3.67
p191	9.19	1.83
p231	12.05	2.66
p271	9.98	1.93
p311	14.91	2.94
ANT-S		
p39	20.07	7.95
p79	36.29	10.18
p119	14.20	8.31
p159	4.43	3.16
p199	3.91	2.43
p239	6.61	3.55
p279	8.27	3.87
pr319	6.22	3.01

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix H.3

Original ANT scale with Low Information Items Removed

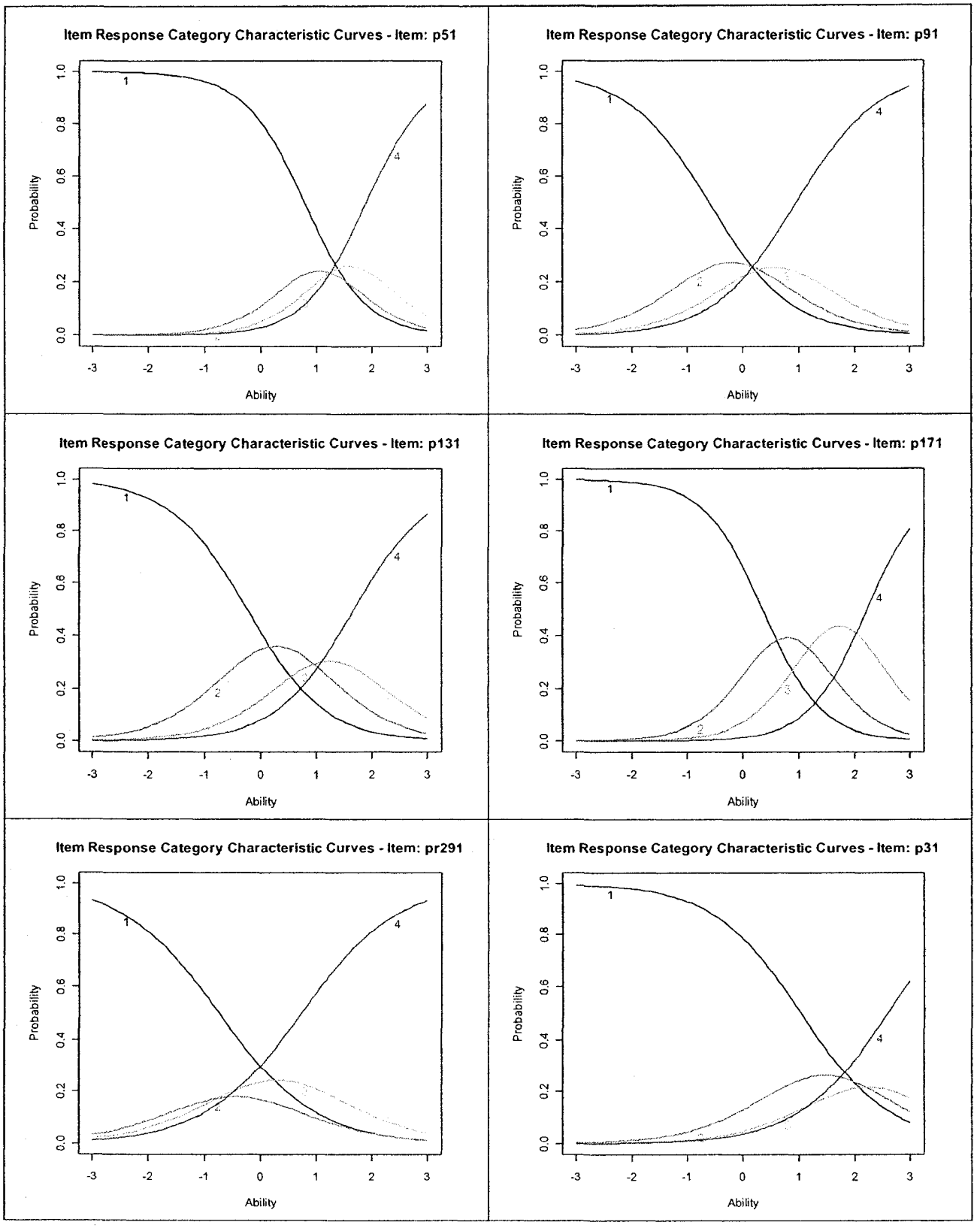

Figure 24. Item Response Category Characteristic Curves (CCC) for ANT-OR Items (1-6)

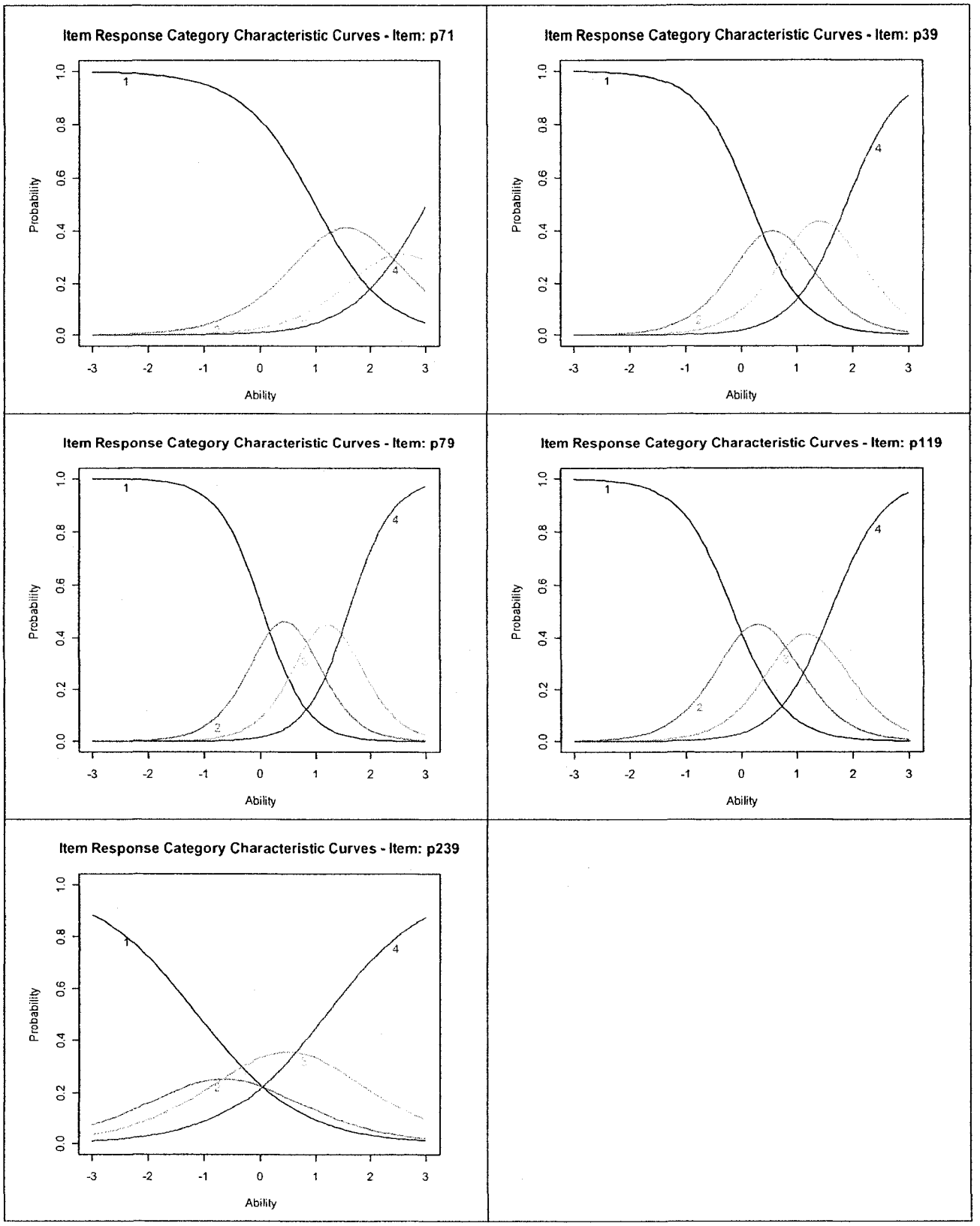

Figure 24 cont'd. Item Response CCCs for ANT-OR Items (7-11)

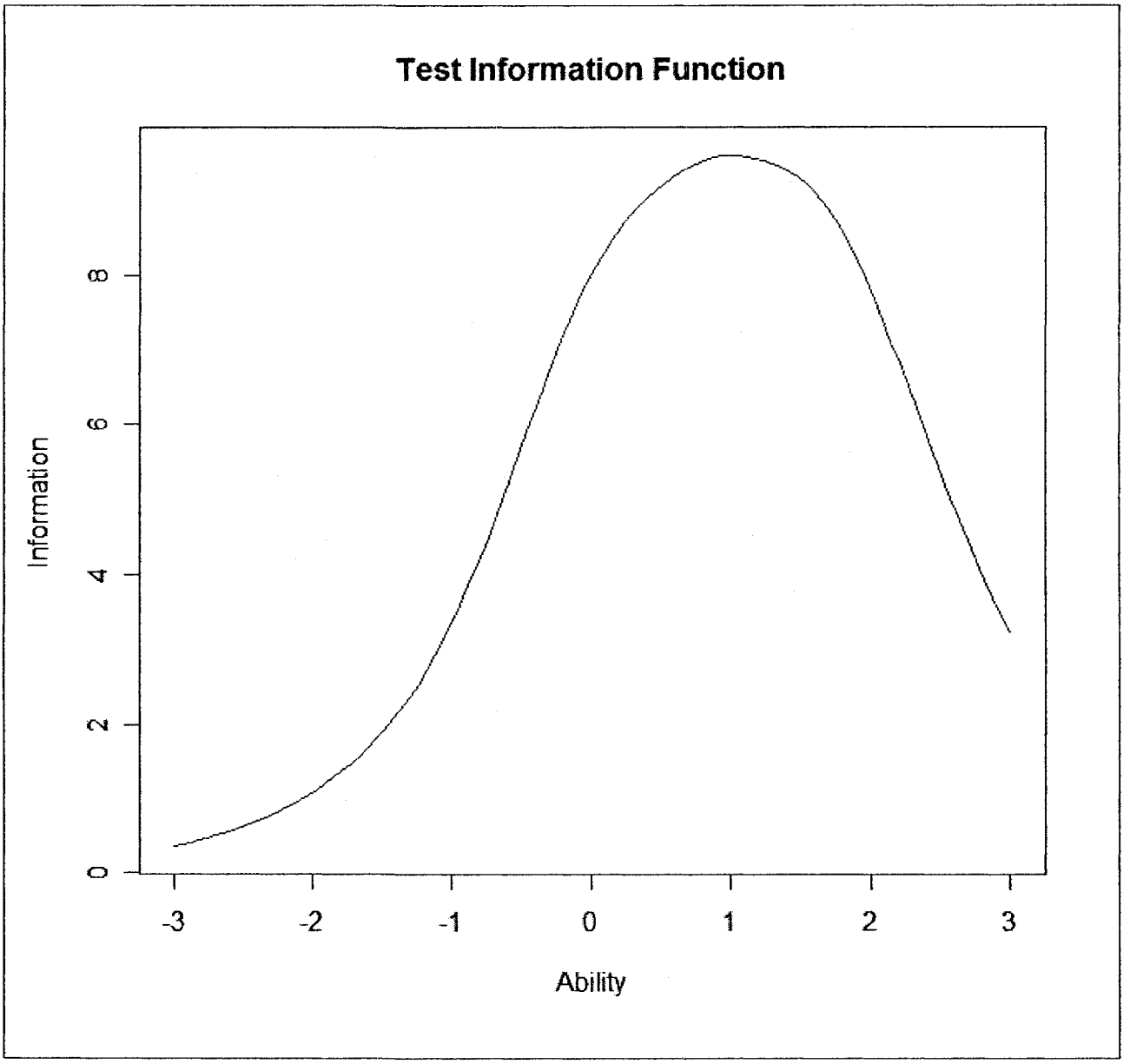

Figure 25. Test Information Function for the original ANT scale with low information items removed.

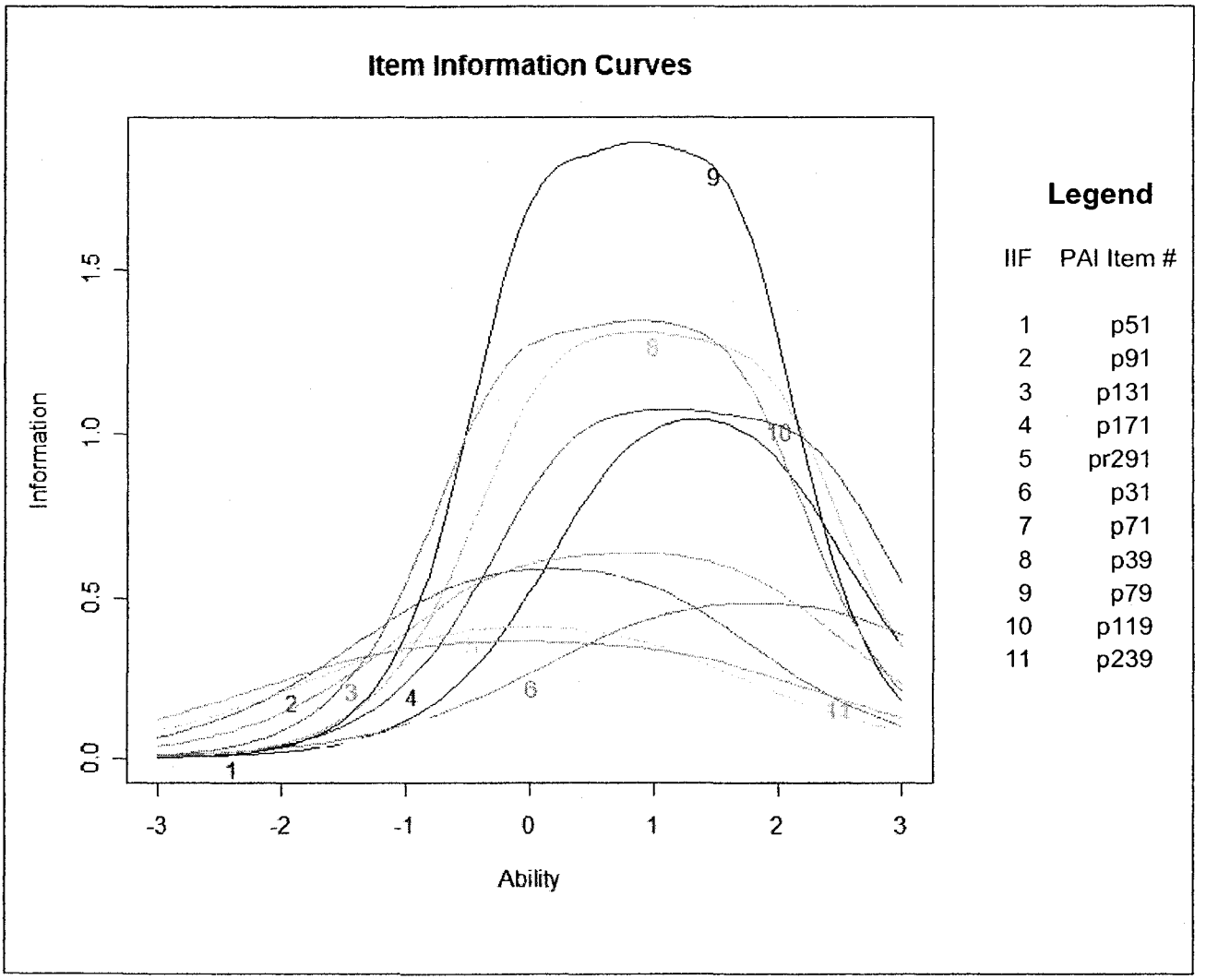

Figure 26. Item Information Functions for the original ANT scale with low information items removed. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 13.1

Modified Original ANT: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p51	0.80	1.33	1.91	1.83	.47
p91	-0.60	0.21	0.97	1.38	.44
p131	-0.24	0.81	1.69	1.44	.43
p171	0.36	1.24	2.23	1.89	.45
pr291	-0.74	-0.12	0.74	1.16	.40
p31	1.04	1.90	2.60	1.25	.41
p71	1.00	2.16	3.02	1.51	.45
p39	0.18	0.99	1.89	2.08	.44
p79	0.05	0.85	1.62	2.53	.47
p119	-0.15	0.78	1.61	2.12	.46
p239	-1.12	-0.17	1.20	1.09	.33
Mean	0.05	0.91	1.77	1.66	*(.43)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 13.2

Modified Original ANT: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	91.4
-2 to +2	73.97
-3 to 0	23.92
0 to +3	67.48
-3 to -2	1.87
-2 to -1	5.76
-1 to 0	16.29
0 to +1	25.94
+1 to +2	25.97
+2 to +3	15.57

Note. [1] = ANT trait range in *SD* units, *M* = 0, *SD* = 1; % = percent of total information or total area under the Test Information Function.

Table 13.3

Modified Original ANT: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p51	8.67
p91	6.67
p131	7.67
p171	11.17
pr291	5.15
p31	5.81
p71	8.36
p39	12.37
p79	15.69
p119	12.80
p239	5.58

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix H.4

IRT results for the newly created ANT scale

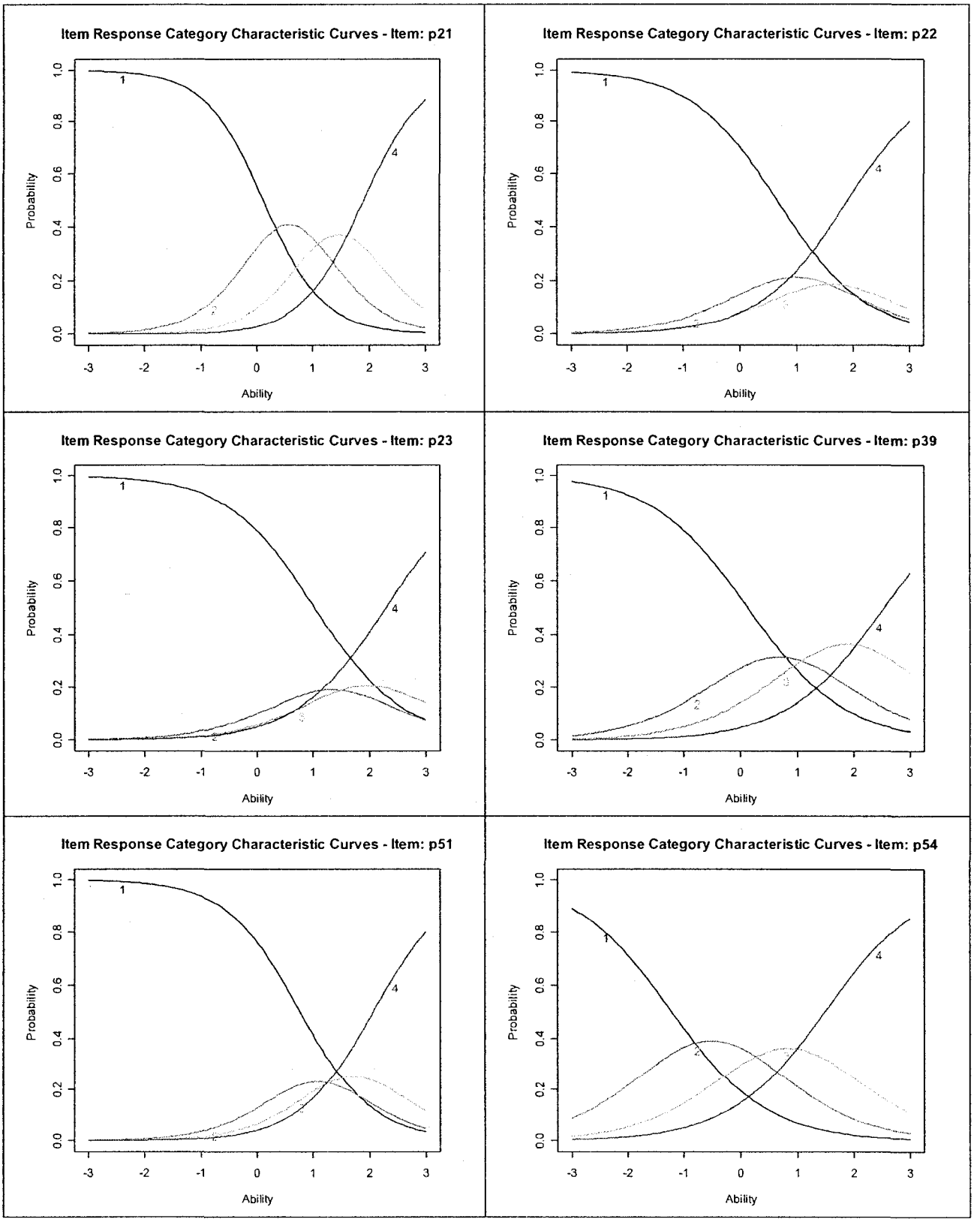

Figure 27. Item Response Category Characteristic Curves (CCC) for ANT-NEW Items (1-6)

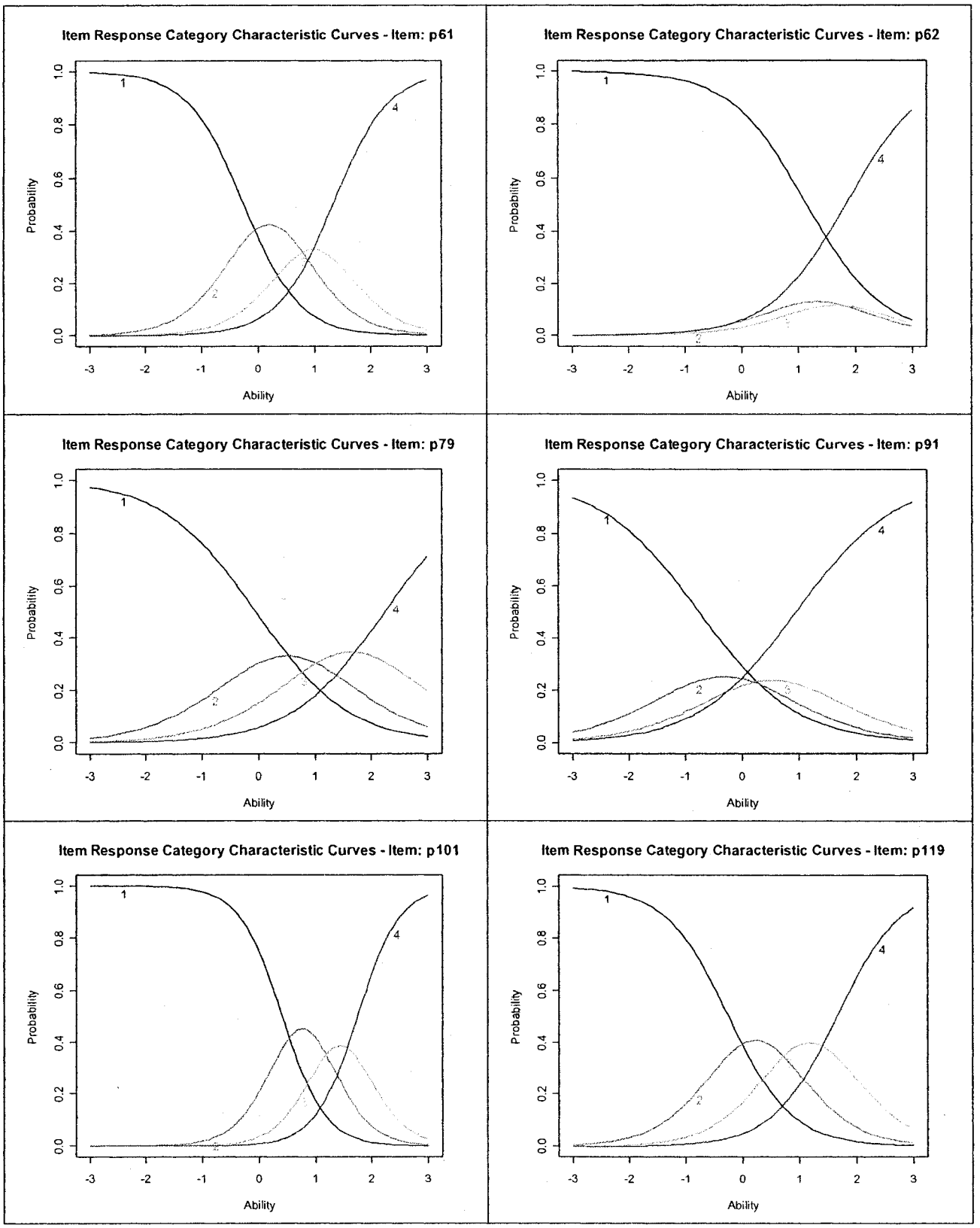

Figure 27 cont'd. Item Response CCCs for ANT-NEW Items (7-12)

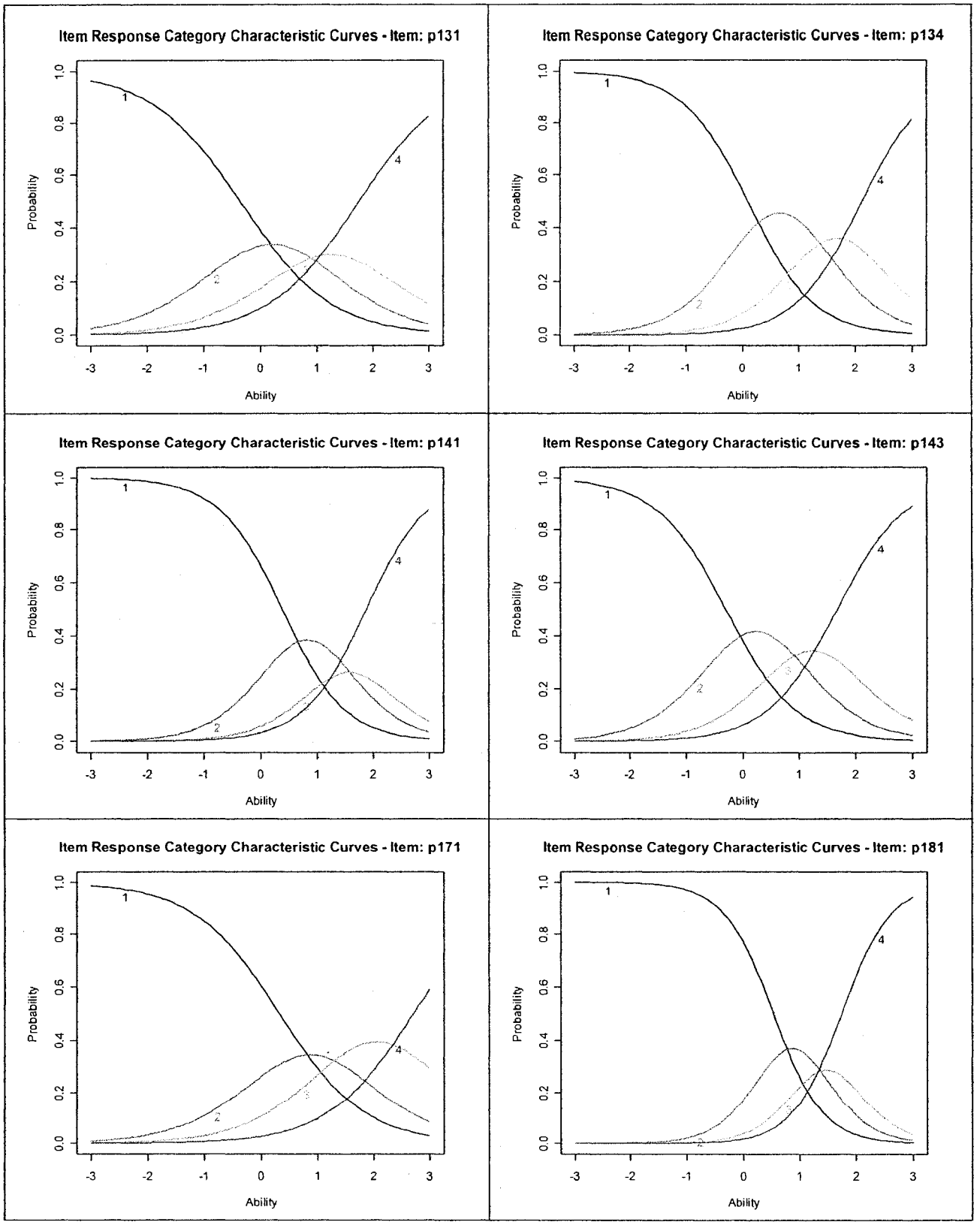

Figure 27 cont'd. Item Response CCCs for ANT-NEW Items (13-18)

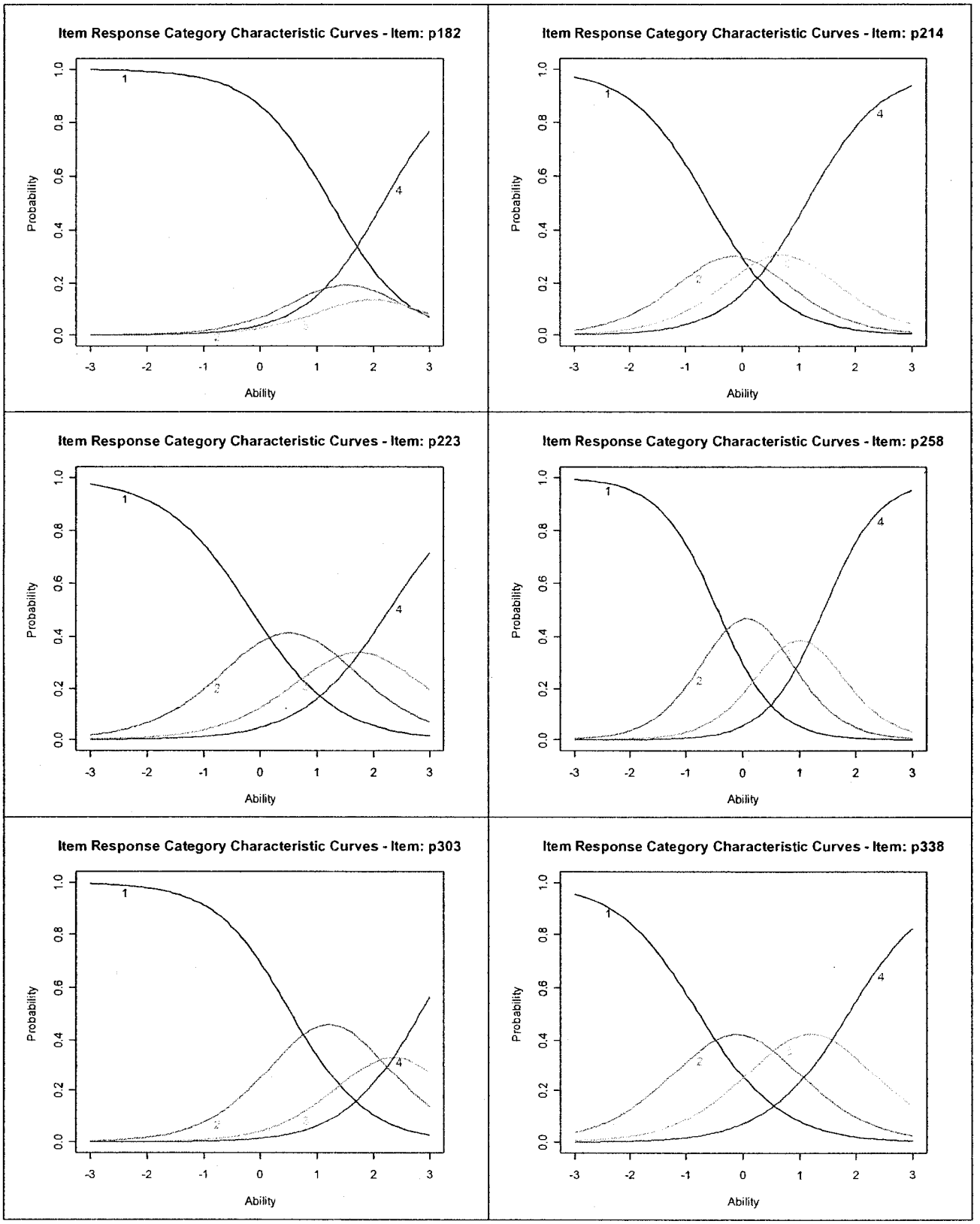

Figure 27 cont'd. Item Response CCCs for ANT-NEW Items (19-24)

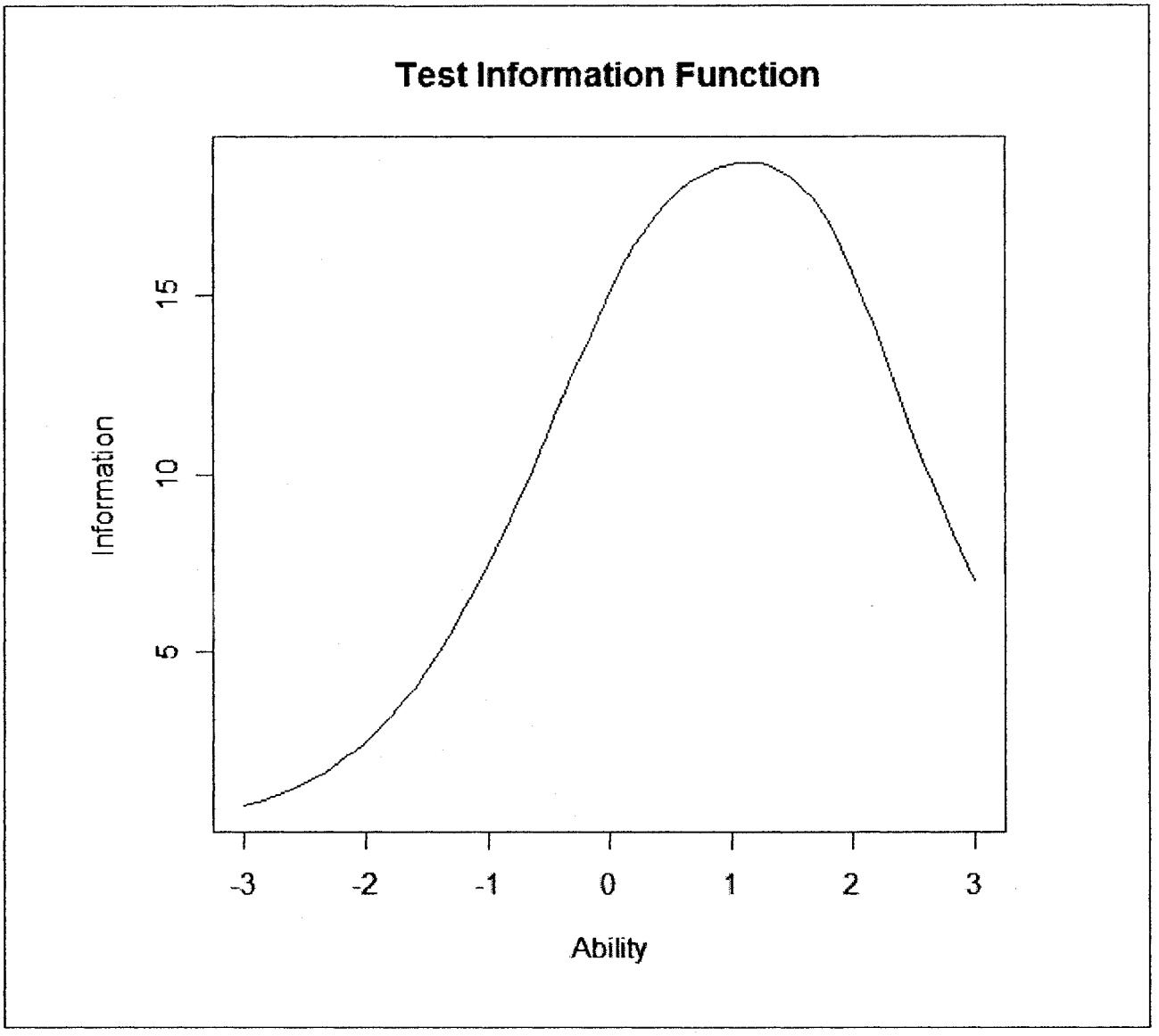

Figure 28. Test Information Function for the new ANT scale.

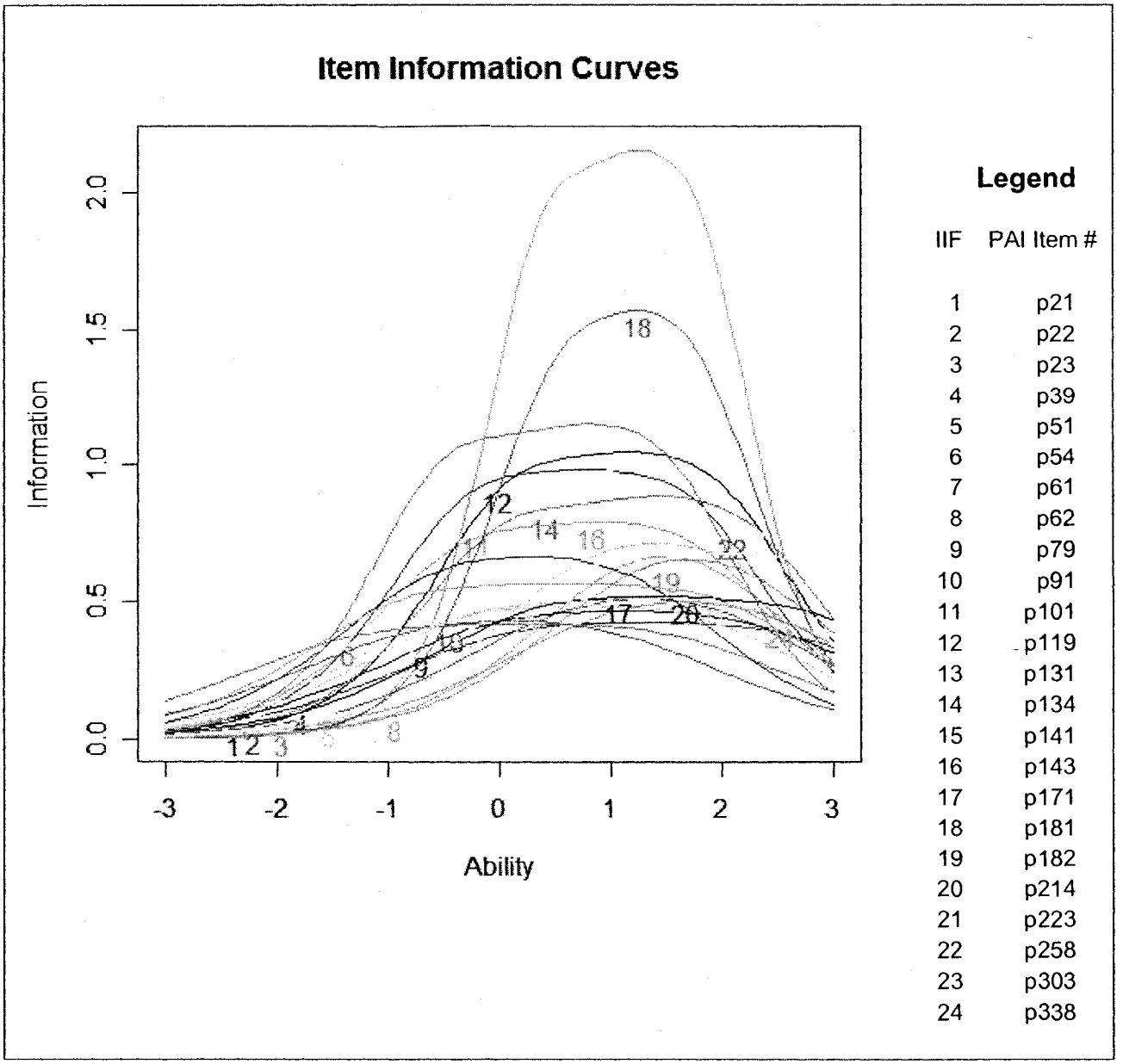

Figure 29. Item Information Functions for the new ANT scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

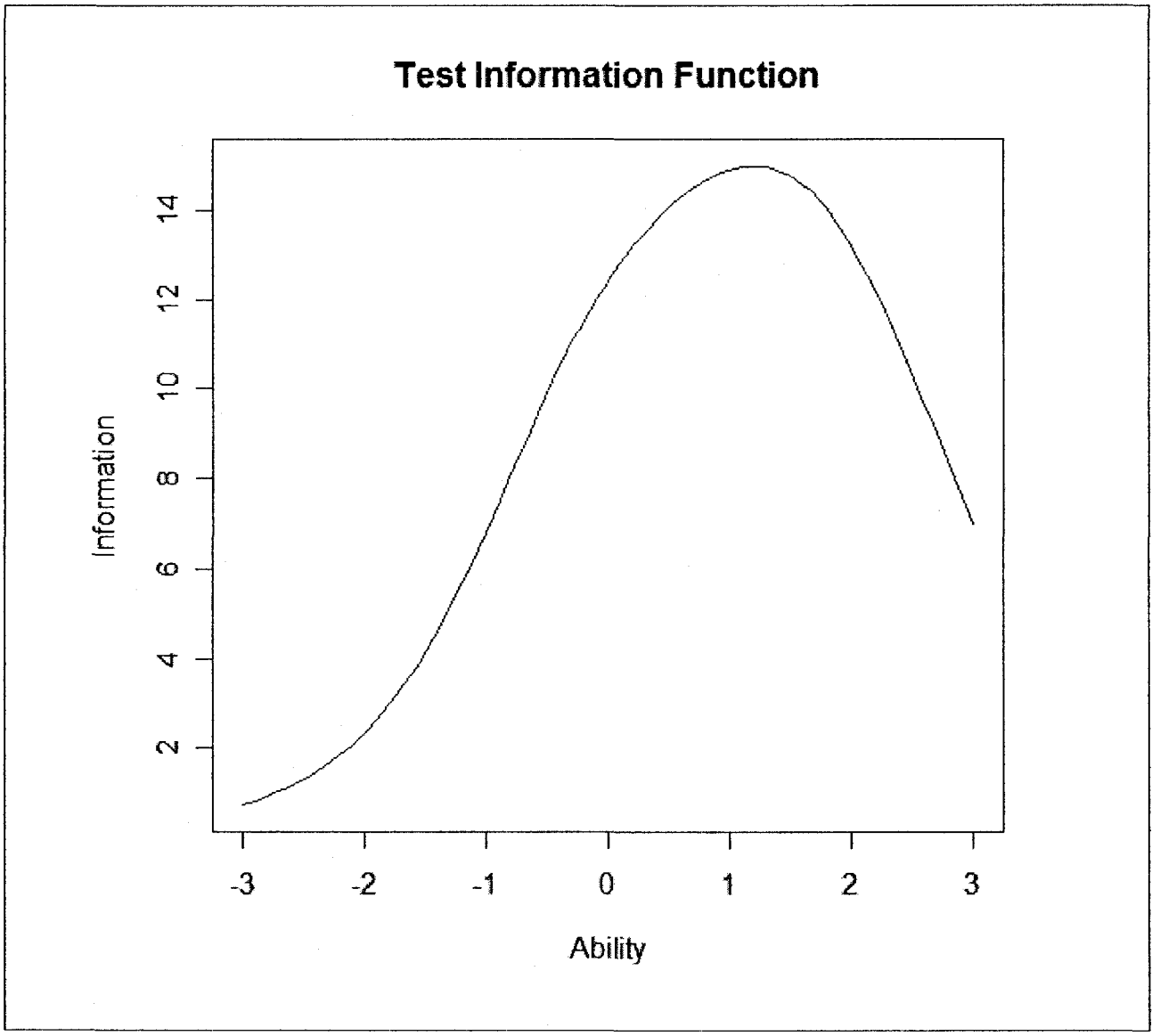

Figure 30. Test Information Function for the new ANT scale with items 101 and 181 removed.

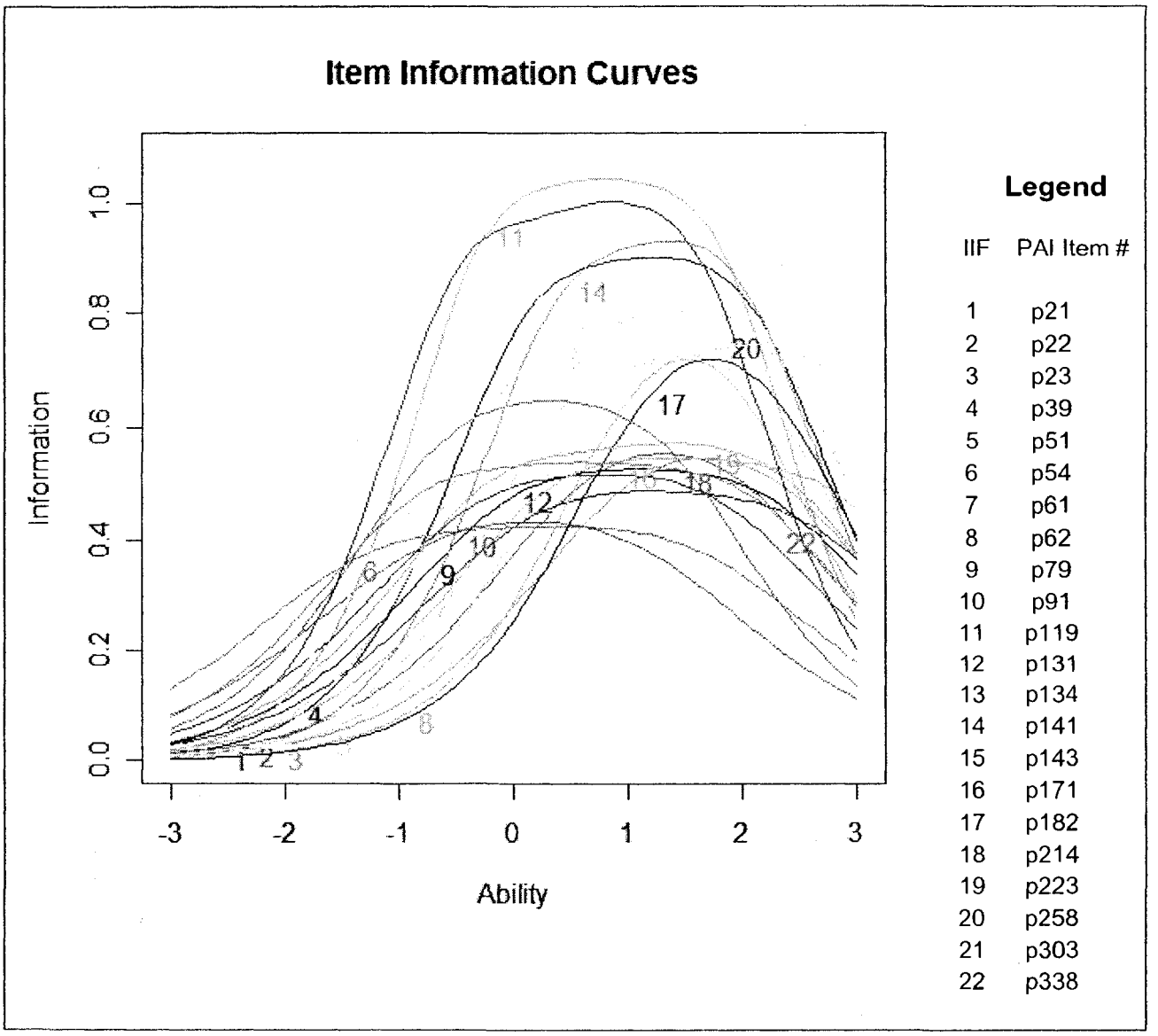

Figure 31. Item Information Functions for the new ANT scale with items 101 and 181 removed. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 14.1

New ANT: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p21	0.12	1.06	1.91	1.86	.45
p22	0.66	1.33	1.91	1.29	.42
p23	1.04	1.64	2.30	1.27	.40
p39	0.13	1.24	2.54	1.18	.38
p51	0.78	1.40	2.07	1.51	.44
p54	-1.22	0.20	1.49	1.17	.39
p61	-0.25	0.65	1.33	2.02	.47
p62	1.16	1.51	1.83	1.50	.44
p79	-0.04	1.08	2.25	1.23	.40
p91	-0.76	0.12	0.95	1.18	.41
p101	0.41	1.14	1.76	2.67	.49
p119	-0.25	0.72	1.66	1.80	.47
p131	-0.36	0.77	1.76	1.25	.41
p134	0.09	1.25	2.13	1.71	.43
p141	0.38	1.28	1.89	1.79	.46
p143	-0.31	0.79	1.69	1.61	.46
p171	0.33	1.43	2.72	1.30	.39
p181	0.54	1.23	1.75	2.25	.48
p182	1.26	1.79	2.17	1.47	.45
p214	-0.59	0.27	1.14	1.46	.41
p223	-0.16	1.20	2.28	1.30	.38
p258	-0.44	0.60	1.43	1.96	.46
p303	0.56	1.89	2.82	1.49	.42
p338	-0.77	0.55	1.87	1.38	.39
Mean	0.10	1.05	1.90	1.57	*(.43)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 14.2

New ANT: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	90.52
-2 to +2	72.56
-3 to 0	24.61
0 to +3	65.91
-3 to -2	2.06
-2 to -1	6.58
-1 to 0	15.97
0 to +1	24.65
+1 to +2	25.36
+2 to +3	15.90

Note. [1] = ANT trait range in *SD* units, *M* = 0, *SD* = 1; % = percent of total information or total area under the Test Information Function.

Table 14.3

New ANT: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p21	5.30
p22	2.80
p23	2.75
p39	3.14
p51	3.51
p54	3.29
p61	5.68
p62	2.83
p79	3.27
p91	2.75
p101	7.89
p119	5.24
p131	3.27
p134	4.97
p141	4.68
p143	4.53
p171	3.62
p181	5.93
p182	3.00
p214	3.73
p223	3.62
p258	5.83
p303	4.29
p338	4.09

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix I

BOR

Appendix I.1

Original BOR Scale

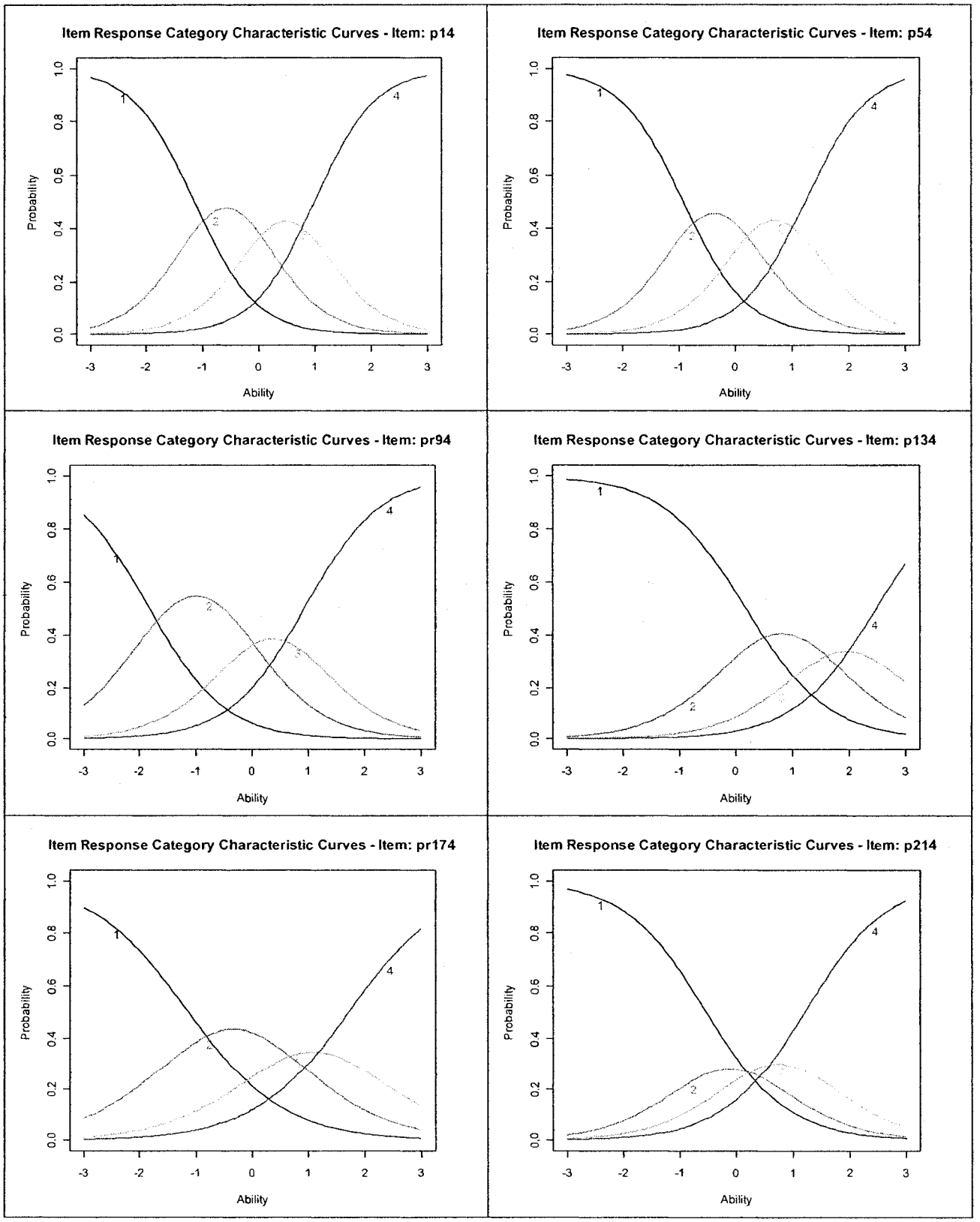

Figure 32. Item Response Category Characteristic Curves (CCC) for BOR Items (1-6)

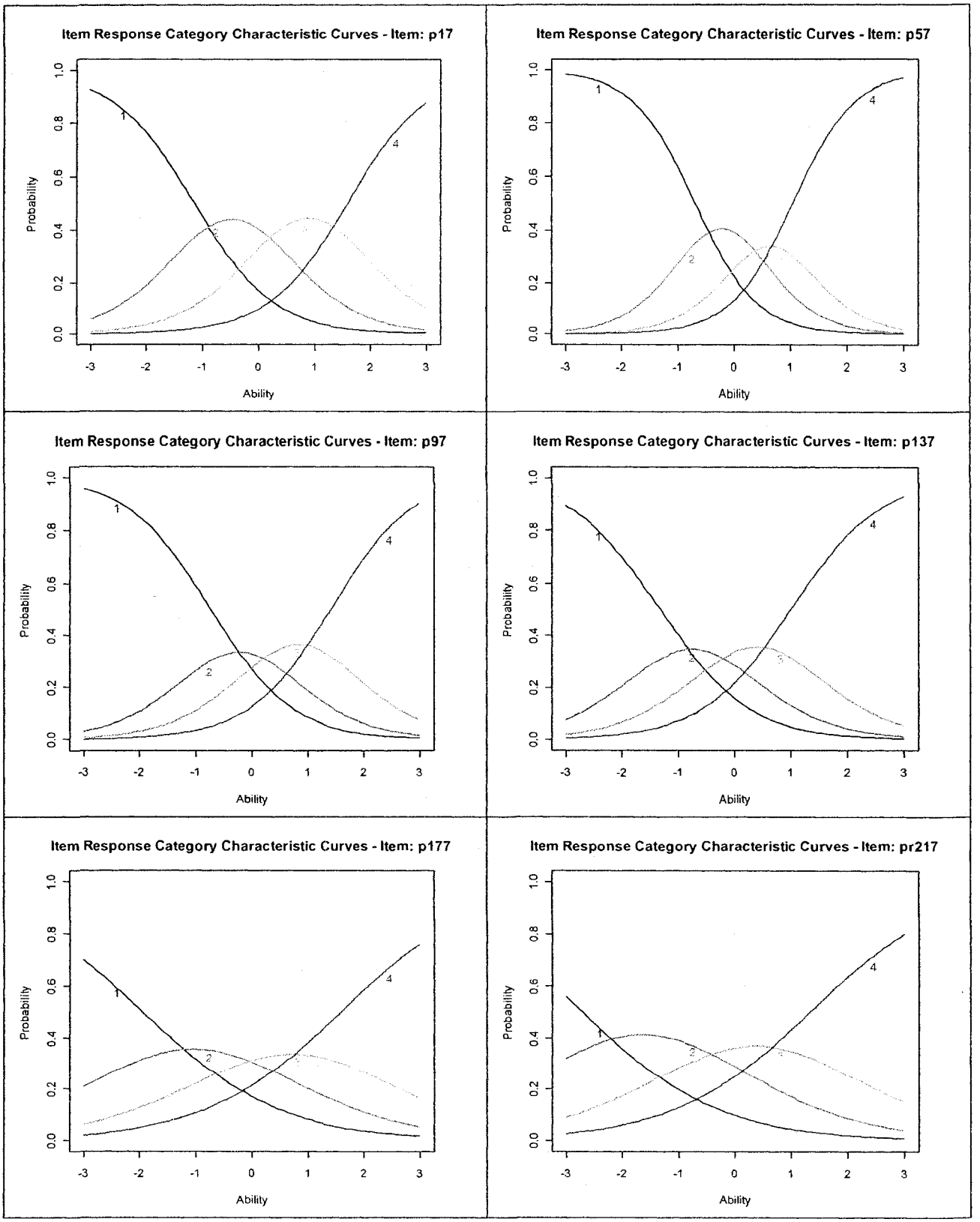

Figure 32 cont'd. Item Response Category Characteristic Curves (CCC) for BOR Items (7-12)

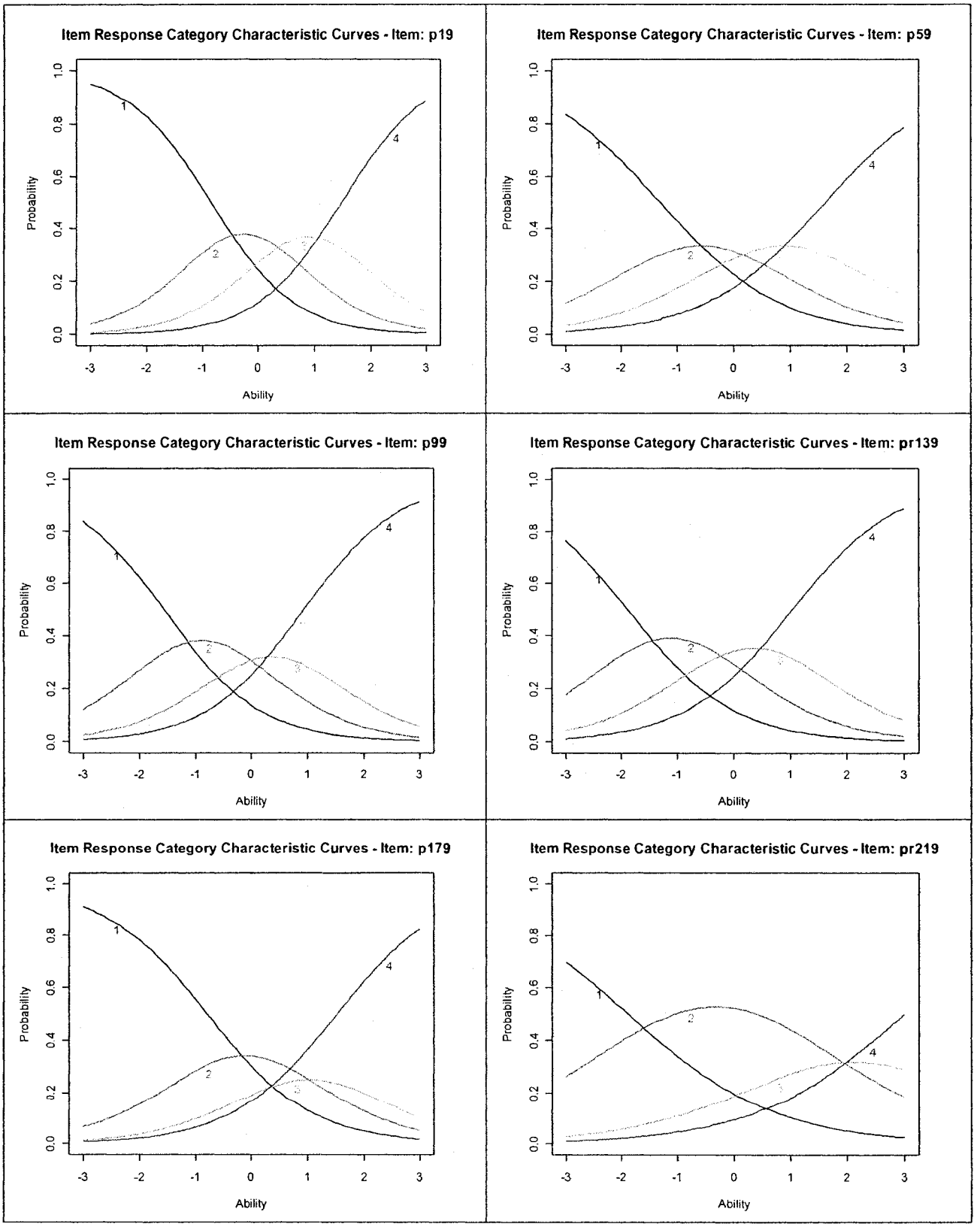

Figure 32 cont'd. Item Response Category Characteristic Curves (CCC) for BOR Items (13-18)

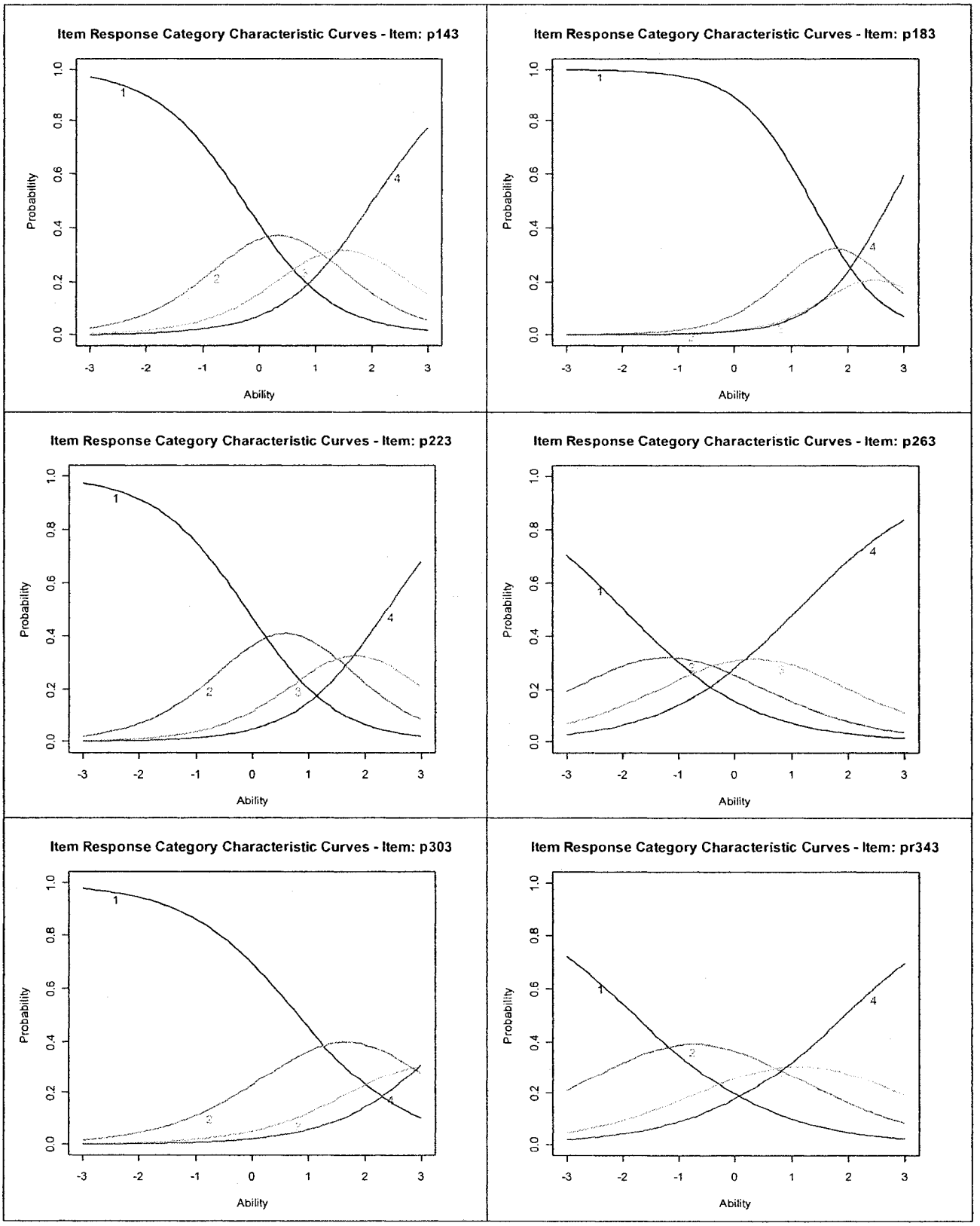

Figure 32 cont'd. Item Response Category Characteristic Curves (CCC) for BOR Items (19-24)

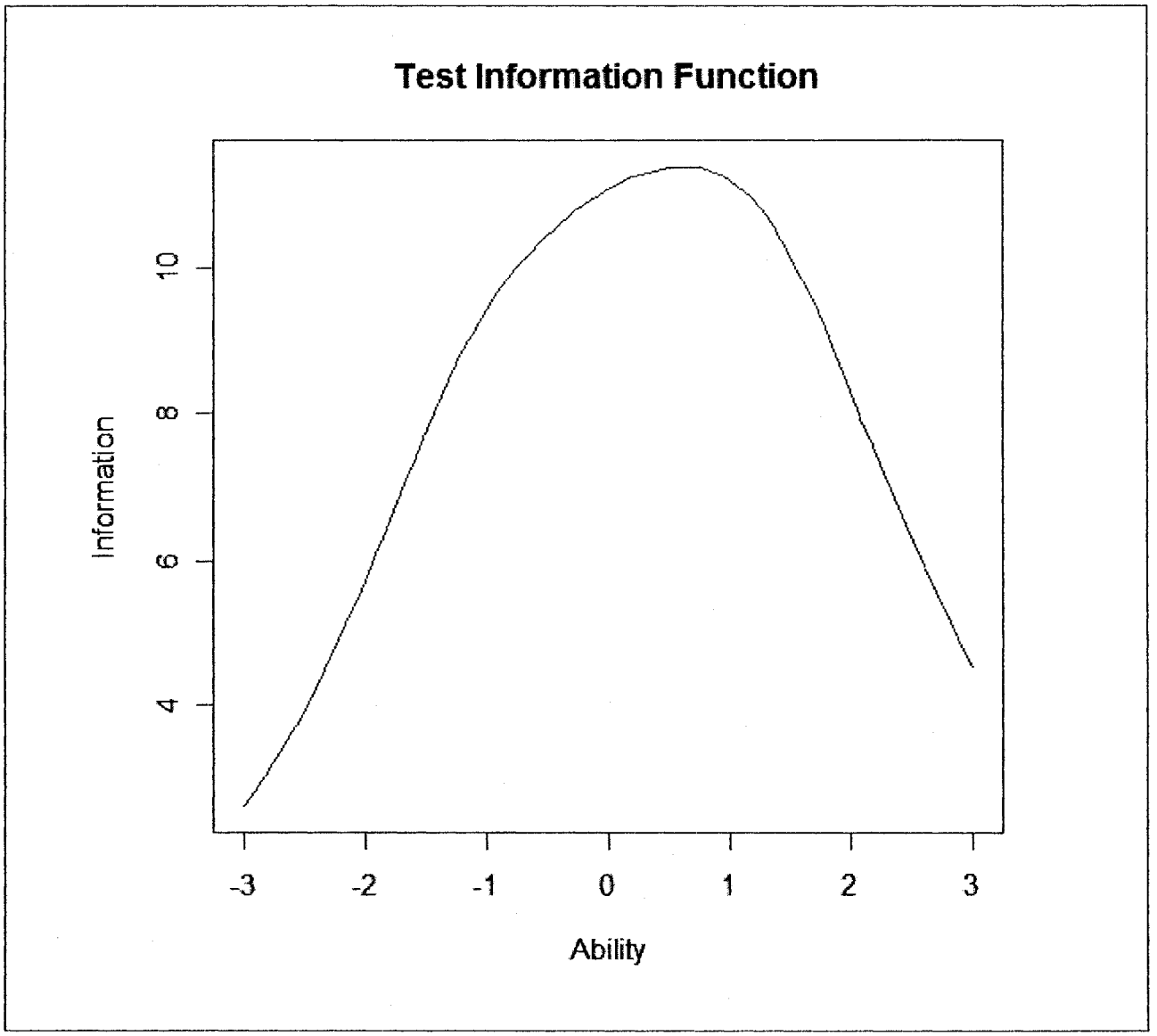

Figure 33. Test Information Function for the original BOR scale.

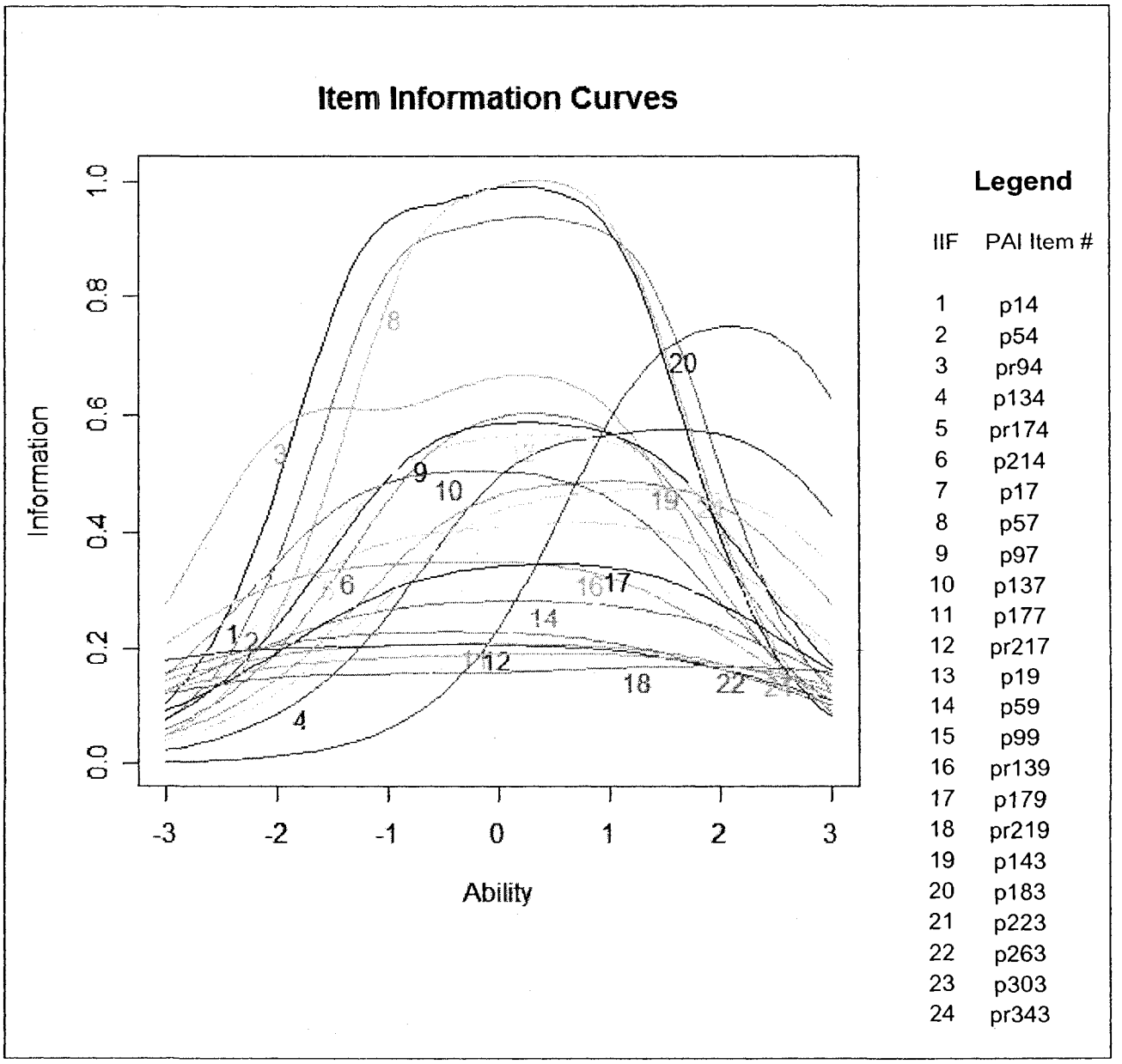

Figure 34. Item Information Functions for the original BOR scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 15.1

Original BOR Scale: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p14	-1.14	-0.01	0.99	1.83	.40
p54	-0.92	0.19	1.23	1.78	.38
pr94	-1.81	-0.17	0.93	1.50	.35
p134	0.19	1.45	2.48	1.37	.37
pr174	-1.13	0.47	1.72	1.17	.35
p214	-0.53	0.31	1.20	1.39	.35
p17	-1.14	0.21	1.59	1.40	.35
p57	-0.70	0.26	1.04	1.81	.40
p97	-0.73	0.29	1.40	1.38	.36
p137	-1.31	-0.18	1.00	1.28	.36
p177	-1.94	-0.13	1.58	0.82	.28
pr217	-2.70	-0.56	1.33	0.83	.28
p19	-0.84	0.34	1.47	1.36	.35
p59	-1.30	0.16	1.62	0.96	.30
p99	-1.57	-0.20	0.94	1.17	.34
pr139	-1.89	-0.34	1.04	1.07	.32
p179	-0.79	0.56	1.53	1.06	.34
pr219	-1.89	1.24	3.01	0.75	.29
p143	-0.29	0.96	2.02	1.26	.37
p183	1.35	2.21	2.76	1.55	.46
p223	-0.10	1.30	2.39	1.24	.36
p263	-1.97	-0.42	1.10	0.86	.31
p303	0.82	2.52	3.79	1.00	.35
pr343	-1.79	0.33	1.94	0.79	.28
Mean	-1.01	0.45	1.67	1.23	*(.34)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 15.2

Original BOR Scale: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	85.65
-2 to +2	67.78
-3 to 0	38.01
0 to +3	47.65
-3 to -2	6.92
-2 to -1	13.21
-1 to 0	17.87
0 to +1	19.44
+1 to +2	17.26
+2 to +3	10.95

Note. [1] = BOR trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 15.3

Original BOR Scale: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p14	6.77
p54	6.53
pr94	5.63
p134	4.63
pr174	4.07
p214	4.22
p17	5.15
p57	6.13
p97	4.55
p137	4.22
p177	2.70
pr217	2.86
p19	4.62
p59	3.08
p99	3.86
pr139	3.63
p179	3.24
pr219	2.69
p143	4.12
p183	4.55
p223	4.19
p263	2.70
p303	3.29
pr343	2.60

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix I.2

Original BOR Subscales

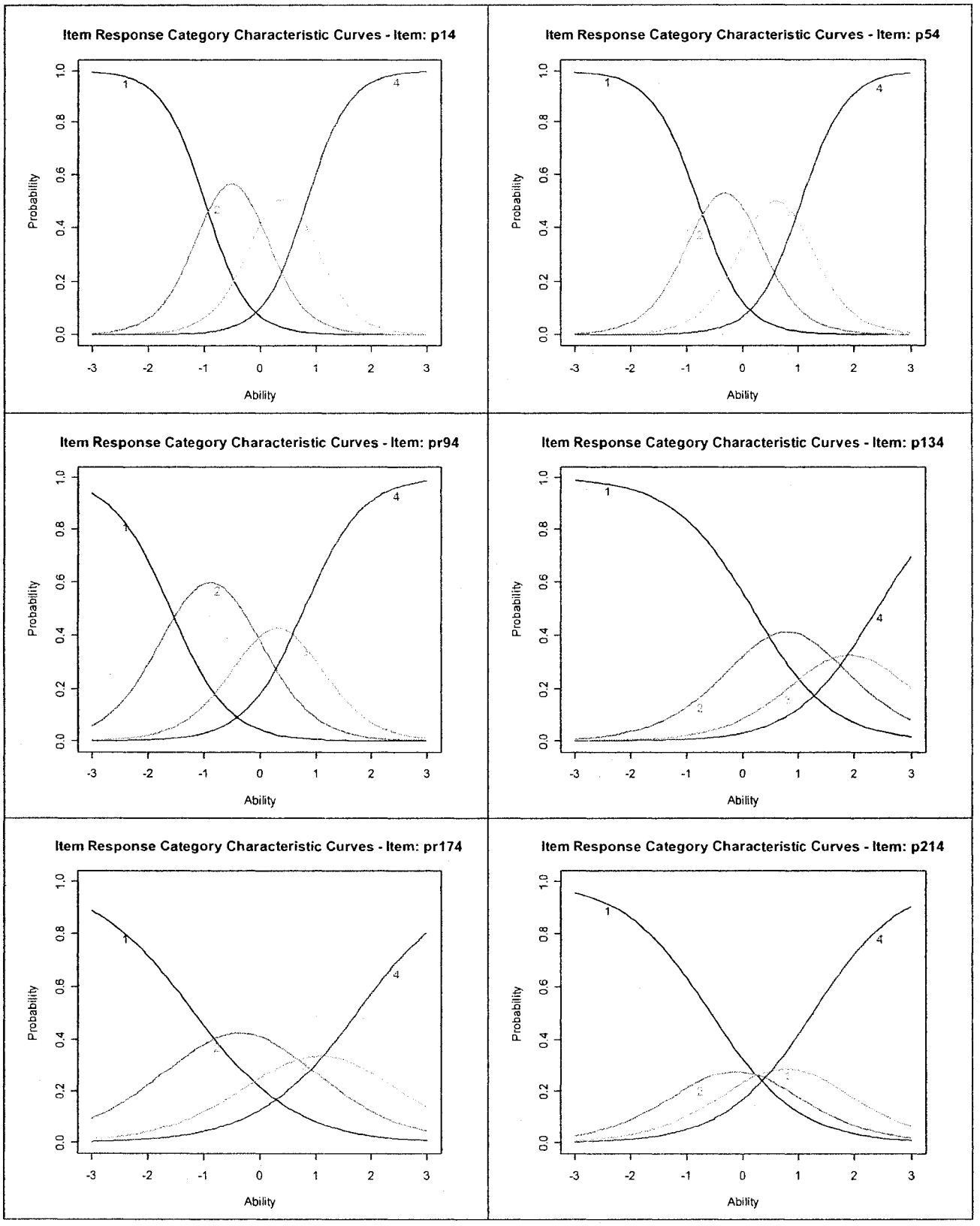

Figure 35. Item Response Category Characteristic Curves (CCC) for BOR-A Items (1-6)

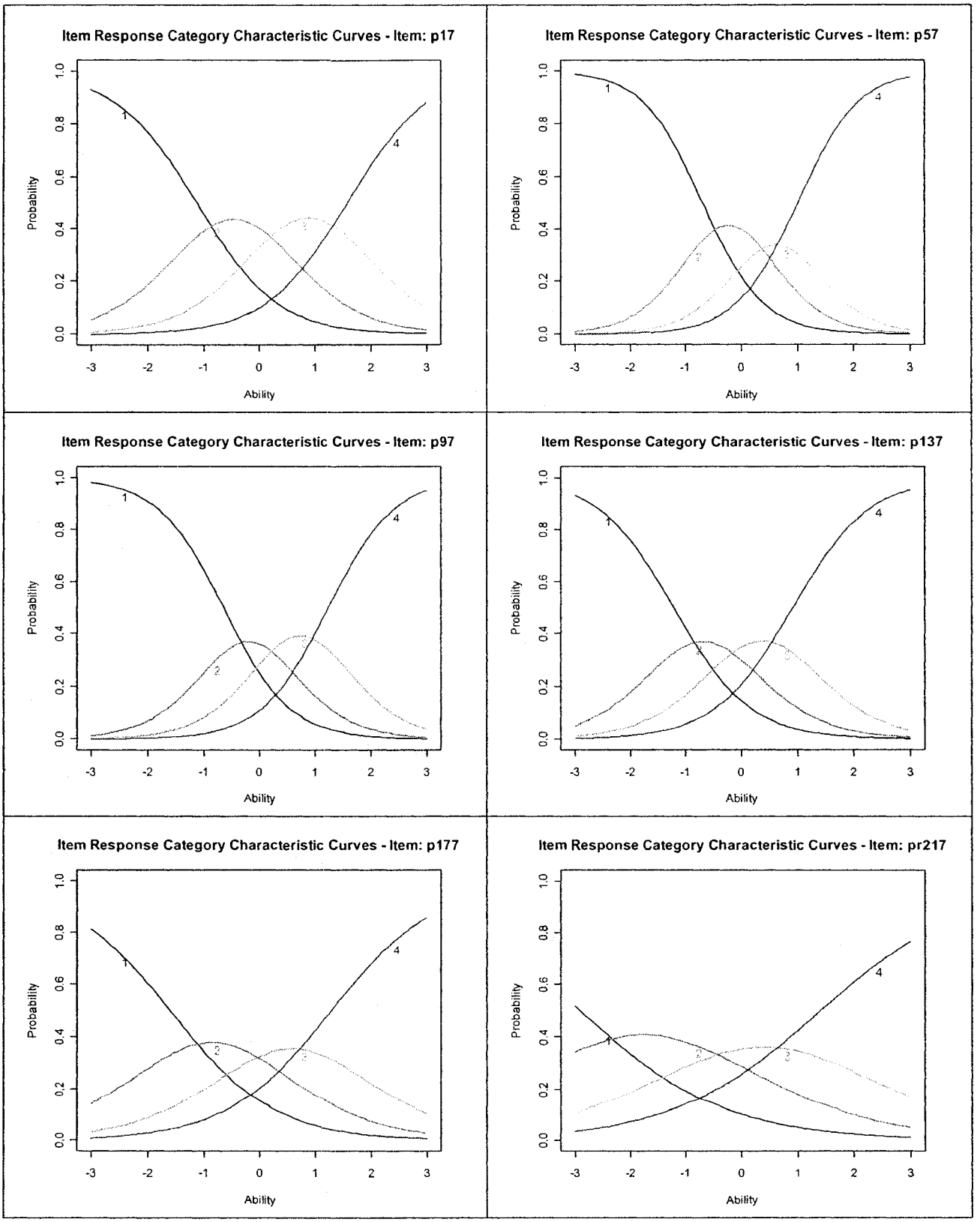

Figure 35 cont'd. Item Response Category Characteristic Curves (CCC) for BOR-I Items (1-6)

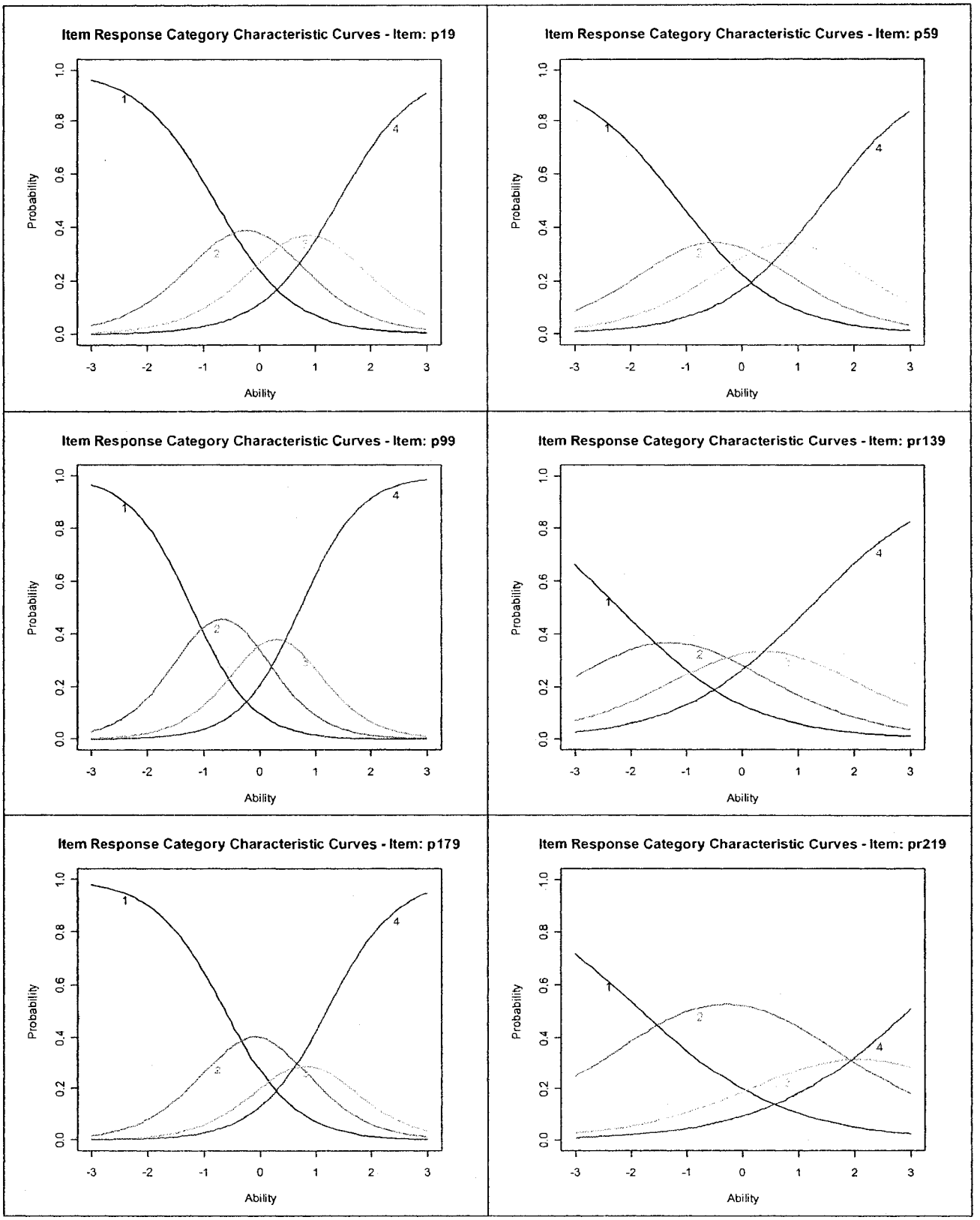

Figure 35 cont'd. Item Response Category Characteristic Curves (CCC) for BOR-N Items (1-6)

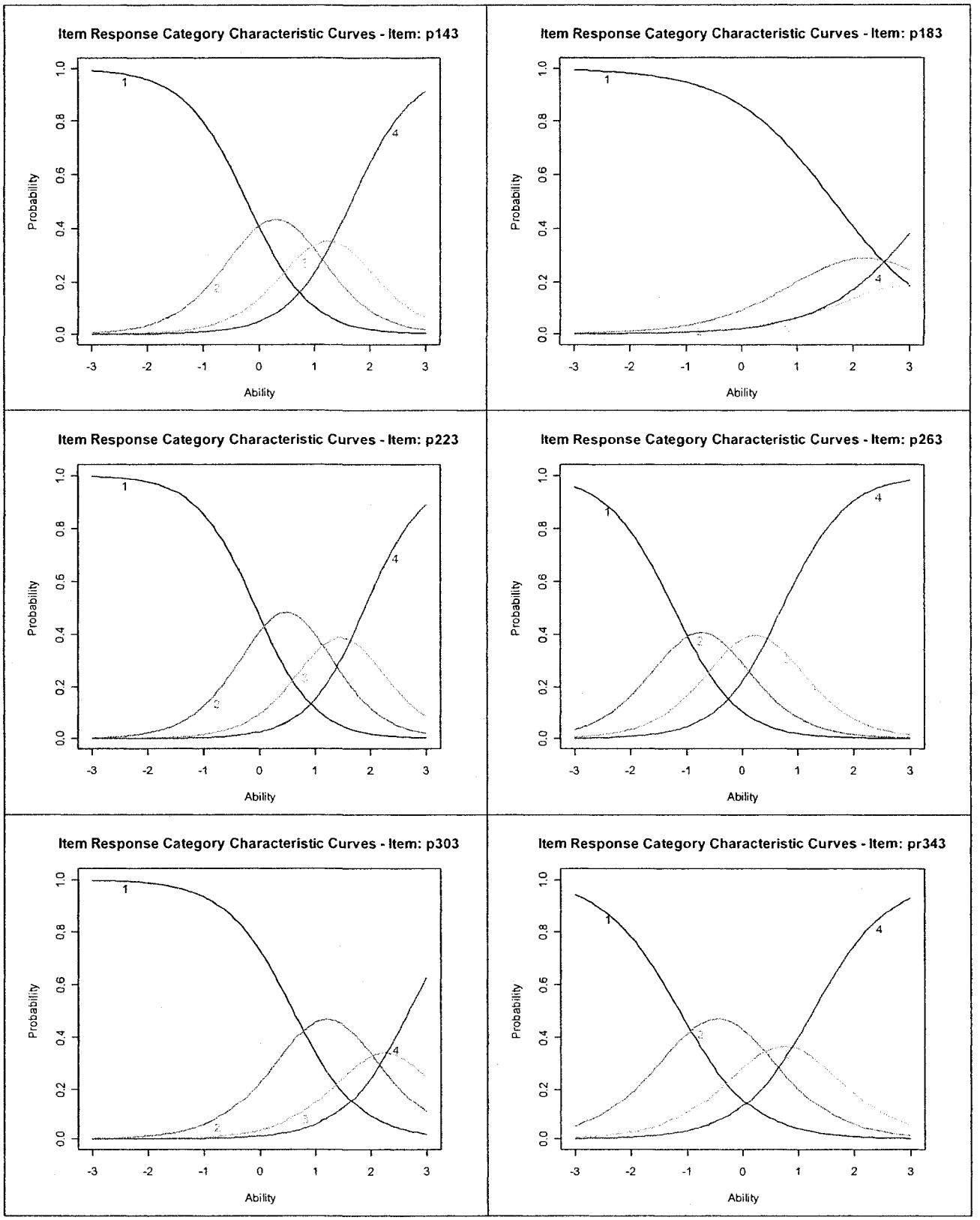

Figure 35 cont'd. Item Response Category Characteristic Curves (CCC) for BOR-S Items (1-6)

Table 15.4

Legend for Figure 36. Corresponding PAI items for the IIF's of the original BOR subscales.

Subscale IIF#[a]	Total Scale IIF#	PAI Items and Original Item Numbers
BOR-A		
1	(1)	14. My mood can shift quite suddenly.
2	(2)	54. My moods get quite intense.
3	(3)	94. My mood is very steady.
4	(4)	134. I have little control over my anger.
5	(5)	174. I've always been a pretty happy person.
6	(6)	214. I've had times when I was so mad I couldn't do enough to express all my anger.
BOR-I		
1	(7)	17. My attitude about myself changes a lot.
2	(8)	57. Sometimes I feel terribly empty inside.
3	(9)	97. I worry a lot about other people leaving me.
4	(10)	137. I often wonder what I should do with my life.
5	(11)	177. I can't handle separation from those close to me very well.
6	(12)	217. I don't get bored very easily.
BOR-N		
1	(13)	19. My relationships have been stormy.
2	(14)	59. I want to let certain people know how much they've hurt me.
3	(15)	99. People once close to me have let me down.
4	(16)	139. I rarely feel very lonely.
5	(17)	179. I've made some real mistakes in the people I've picked as friends.
6	(18)	219. Once someone is my friend, we stay friends.
BOR-S		
1	(19)	143. I sometimes do things so impulsively that I get into trouble.
2	(20)	183. When I'm upset, I typically do something to hurt myself.
3	(21)	223. I'm too impulsive for my own good.
4	(22)	263. I spend money too easily.
5	(23)	303. I'm a reckless person.
6	(24)	343. I'm careful about how I spend my money.

Note. [a]IIF# = Item Information Function number from Figure 36.

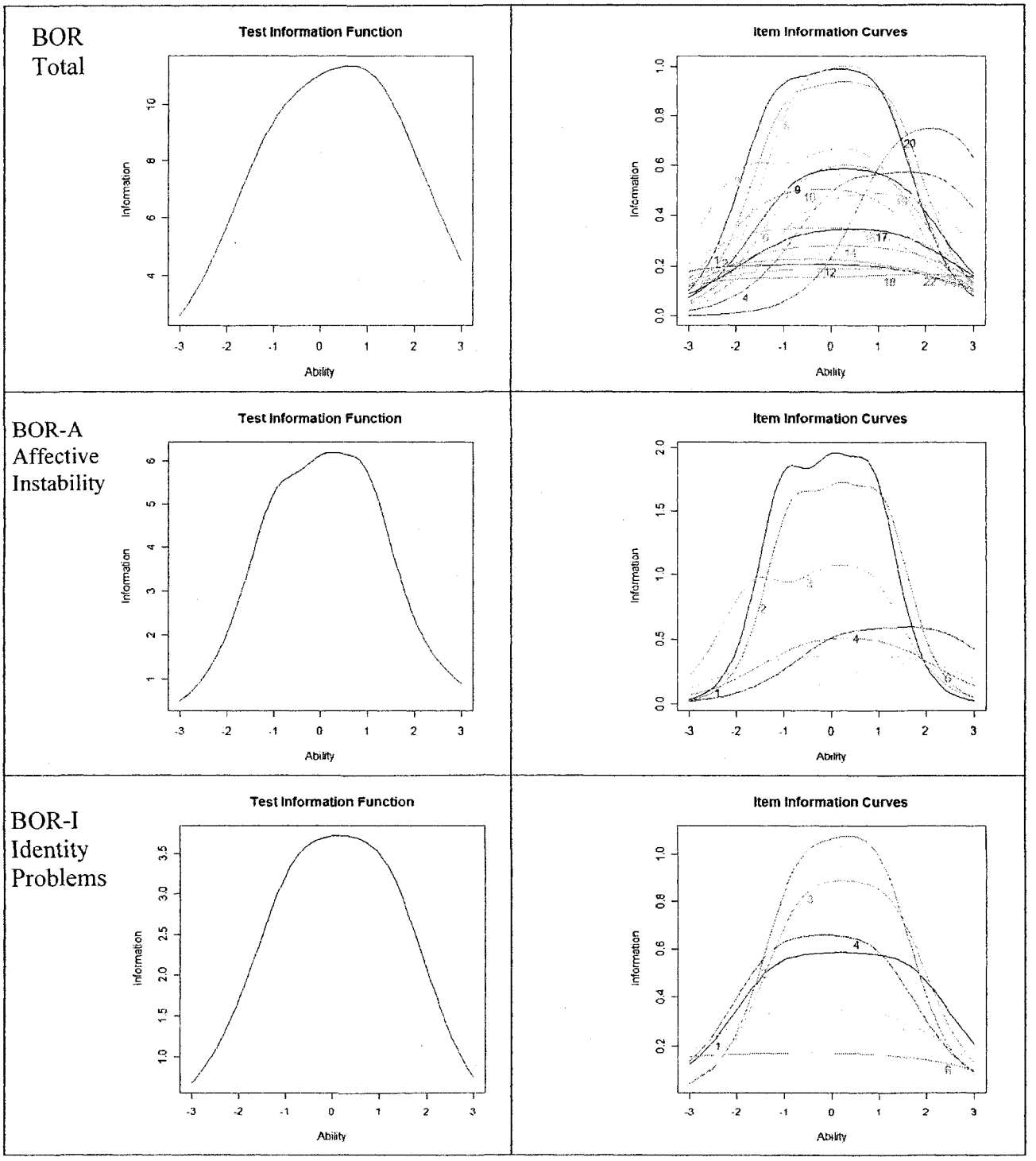

Figure 36. Test and item information functions for the original BOR ANT subscales (see Table 15.4 for a legend of the PAI items that correspond with the IIF numbers).

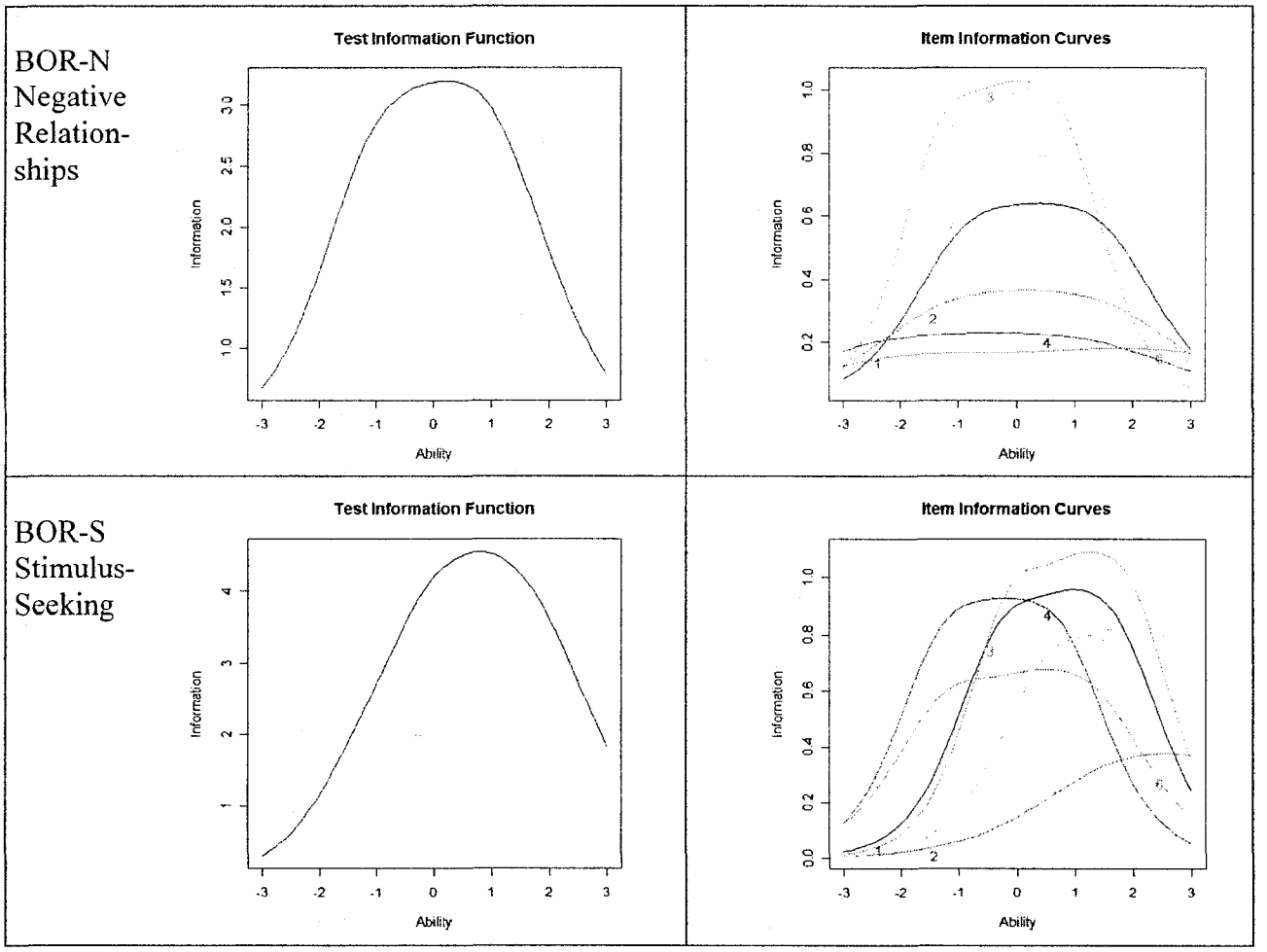

Figure 36 cont'd. Test and item information functions for the original PAI BOR subscales (see Table 15.4 for a legend of the PAI items that correspond with the IIF numbers).

Table 15.5

Original BOR subscales: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	Hg *(H)*
BOR-A					
p14	-0.98	-0.01	0.84	2.62	.50
p54	-0.80	0.17	1.07	2.45	.48
pr94	-1.59	-0.15	0.80	1.92	.45
p134	0.18	1.43	2.40	1.40	.43
pr174	-1.16	0.47	1.73	1.12	.39
p214	-0.57	0.32	1.25	1.28	.39
Mean	*-0.82*	*0.37*	*1.35*	*1.80*	*(.44)*
BOR-I					
p17	-1.12	0.22	1.57	1.40	.36
p57	-0.69	0.25	1.00	1.88	.40
p97	-0.64	0.27	1.24	1.71	.39
p137	-1.21	-0.15	0.91	1.47	.37
p177	-1.60	-0.10	1.29	1.07	.31
pr217	-2.91	-0.60	1.41	0.75	.27
Mean	*-1.36*	*-0.02*	*1.24*	*1.38*	*(.35)*
BOR-N					
p19	-0.80	0.34	1.42	1.45	.35
p59	-1.15	0.17	1.49	1.08	.30
p99	-1.21	-0.13	0.73	1.85	.37
pr139	-2.20	-0.41	1.20	0.86	.29
p179	-0.62	0.45	1.20	1.60	.38
pr219	-1.80	1.23	2.95	0.77	.28
Mean	*-1.29*	*0.28*	*1.50*	*1.27*	*(.33)*
BOR-S					
p143	-0.22	0.83	1.66	1.77	.46
p183	1.65	2.73	3.43	1.10	.43
p223	-0.06	1.04	1.90	1.91	.43
p263	-1.23	-0.24	0.73	1.75	.49
p303	0.60	1.83	2.69	1.66	.45
pr343	-1.13	0.24	1.27	1.50	.43
Mean	*-0.06*	*1.07*	*1.95*	*1.61*	*(.45)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 15.6

Original BOR Subscales: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information				
	Total Scale	BOR-A	BOR-I	BOR-N	BOR-S
-3 to +3	85.65	94.99	90.99	88.34	89.51
-2 to +2	67.78	83.85	76.19	72.89	72.14
-3 to 0	38.01	45.02	43.31	42.93	31.13
0 to +3	47.65	49.97	47.68	45.42	58.37
-3 to -2	6.92	4.76	6.76	7.19	3.34
-2 to -1	13.21	15.65	14.97	15.28	9.74
-1 to 0	17.87	24.61	21.58	20.46	18.05
0 to +1	19.44	26.19	22.23	20.97	22.88
+1 to +2	17.26	17.40	17.41	16.18	21.47
+2 to +3	10.95	6.38	8.04	8.26	14.03

Note. [1] = BOR trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 15.7

Original BOR Subscales: Item Information as a Function of Trait Level (Theta)

PAI Items	Percent of Total Information	
	BOR Subscale	BOR Total Scale
BOR-A		
p14	26.28	6.77
p54	24.13	6.53
pr94	18.73	5.63
p134	11.70	4.63
pr174	9.56	4.07
p214	9.56	4.22
BOR-I		
p17	18.09	5.15
p57	22.38	6.13
p97	20.51	4.55
p137	17.42	4.22
p177	12.58	2.70
pr217	9.01	2.86
BOR-N		
p19	19.08	4.62
p59	13.63	3.08
p99	25.53	3.86
pr139	10.90	3.63
p179	20.21	3.24
pr219	10.64	2.69
BOR-S		
p143	18.36	4.12
p183	9.21	4.55
p223	20.73	4.19
p263	18.31	2.70
p303	17.39	3.29
pr343	15.95	2.60

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix I.3

Modified Original BOR: Low Information Items Removed

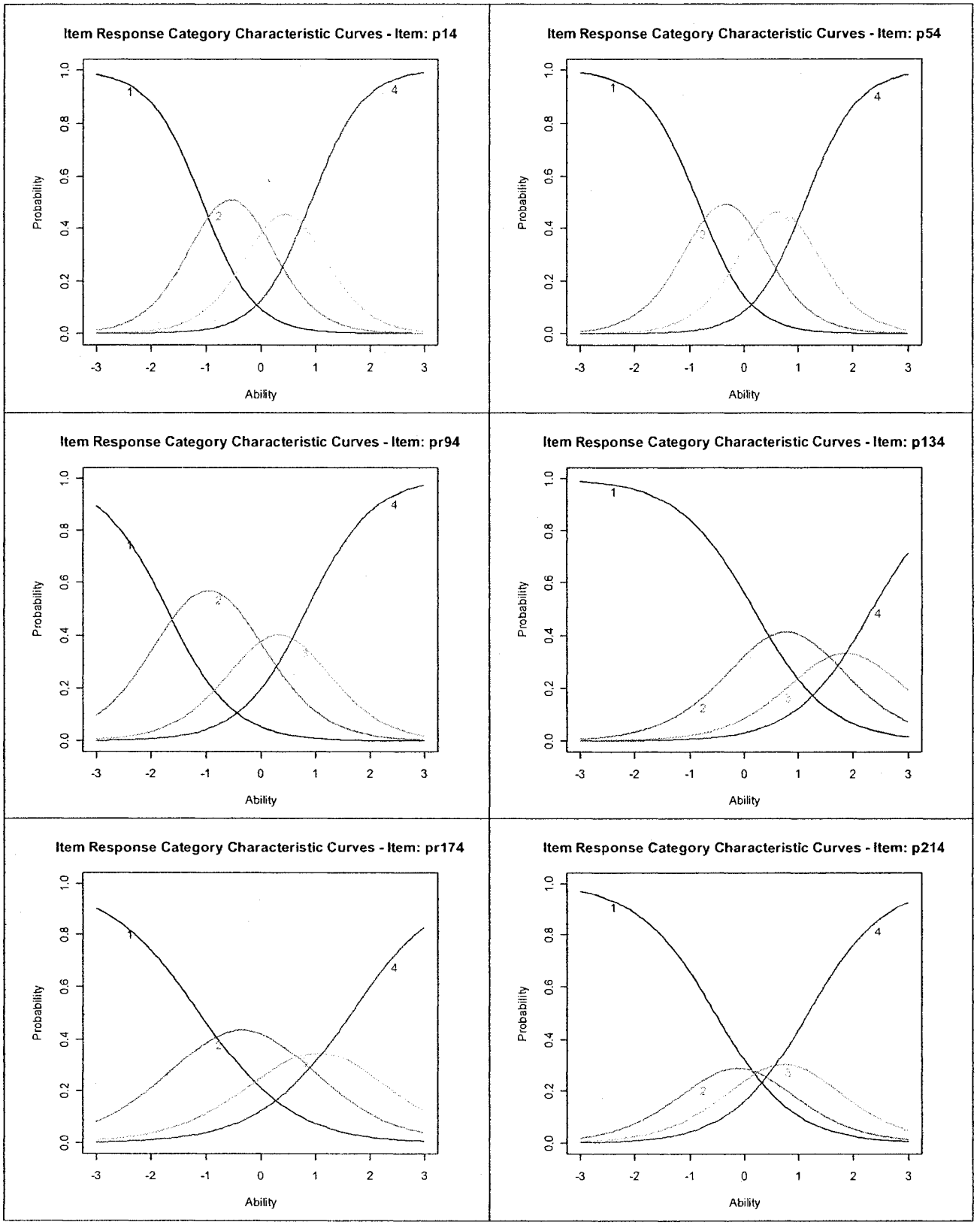

Figure 37. Item Response Category Characteristic Curves (CCC) for BOR-OR Items (1-6)

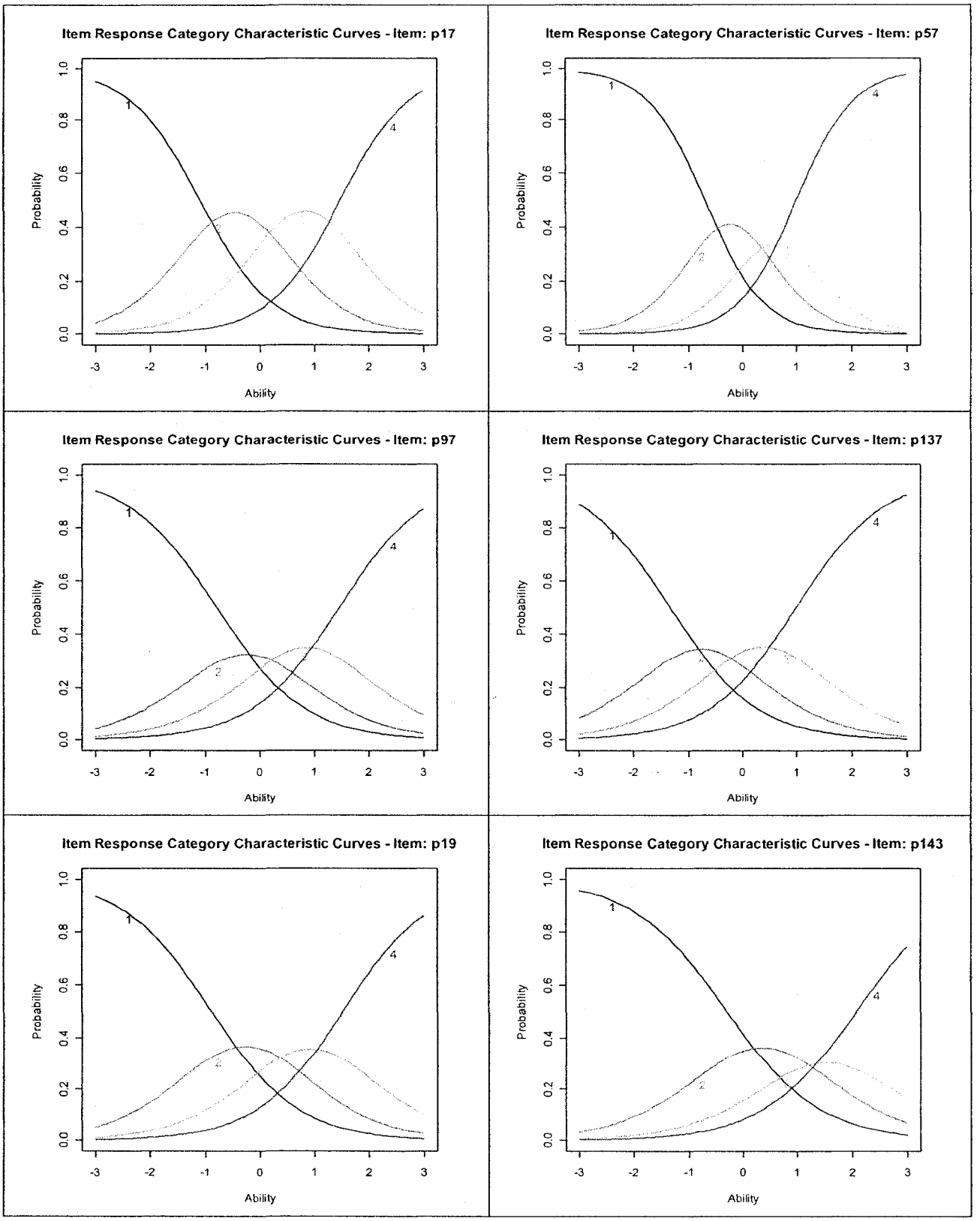

Figure 37 cont'd. Item Response CCCs for BOR-OR Items (7-12)

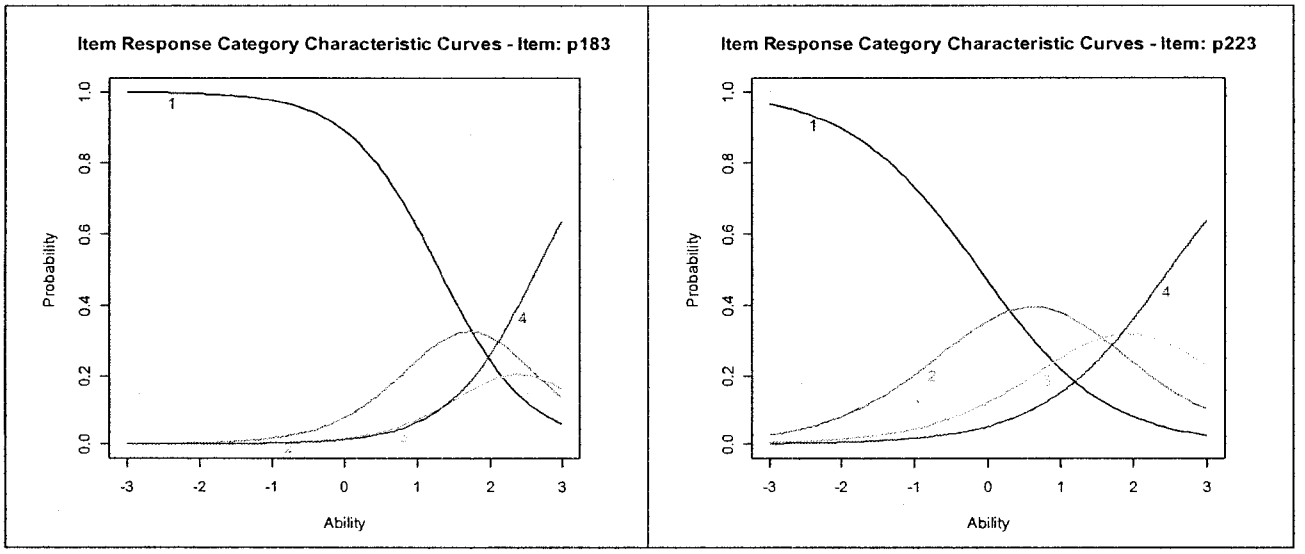

Figure 37 cont'd. Item Response CCCs for BOR-OR Items (13-14)

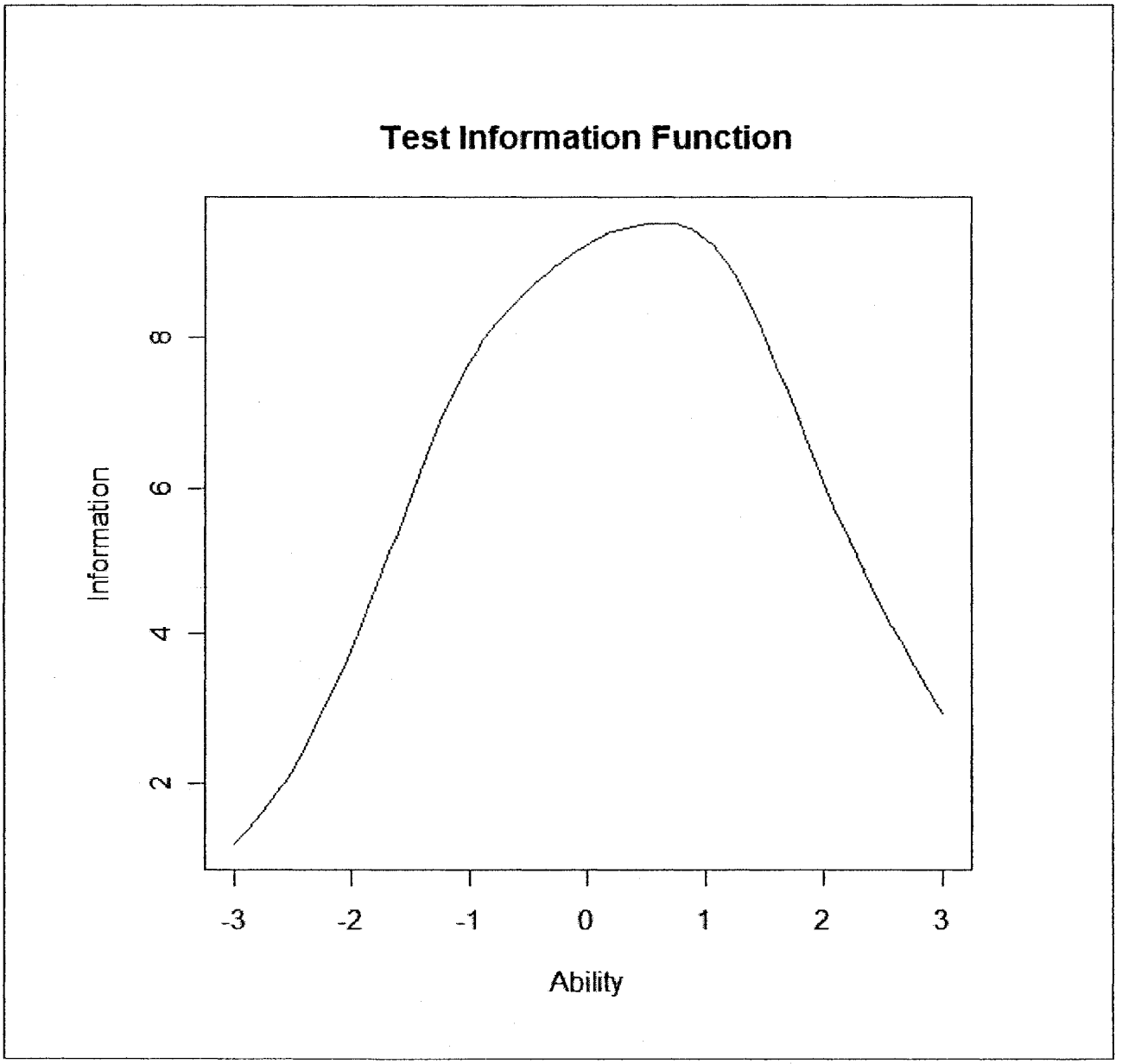

Figure 38. Test Information Function for the original BOR scale with low information items removed.

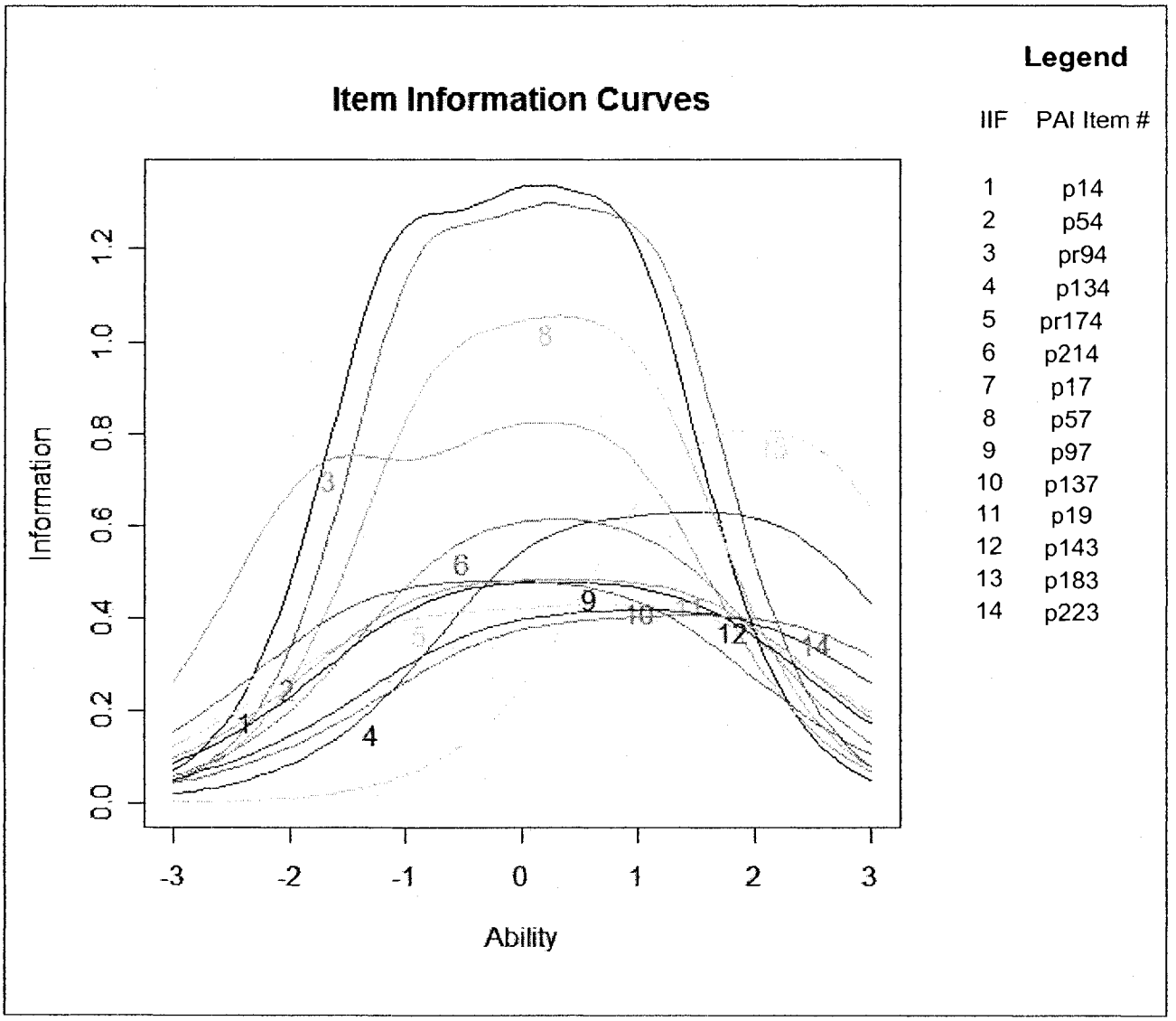

Figure 39. Item Information Functions for the original BOR scale with low information items removed. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 16.1

Modified Original BOR Scale: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	Hg *(H)*
p14	-1.06	-0.01	0.91	2.13	.46
p54	-0.85	0.17	1.13	2.10	.44
pr94	-1.71	-0.17	0.85	1.67	.40
p134	0.17	1.40	2.36	1.44	.41
pr174	-1.12	0.45	1.67	1.19	.38
p214	-0.54	0.30	1.18	1.40	.38
p17	-1.09	0.19	1.48	1.54	.40
p57	-0.70	0.25	1.00	1.86	.45
p97	-0.79	0.30	1.46	1.24	.37
p137	-1.34	-0.19	0.99	1.25	.40
p19	-0.88	0.35	1.53	1.26	.38
p143	-0.30	1.01	2.08	1.16	.39
p183	1.30	2.14	2.65	1.62	.49
p223	-0.11	1.35	2.50	1.15	.37
Mean	-0.64	0.54	1.56	1.50	*(.41)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 16.2

Modified Original BOR Scale: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	91.34
-2 to +2	75.45
-3 to 0	39.63
0 to +3	51.71
-3 to -2	5.44
-2 to -1	13.81
-1 to 0	20.38
0 to +1	22.44
+1 to +2	18.83
+2 to +3	10.45

Note. [1] = BOR trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 16.3

Modified Original BOR Scale: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p14	11.27
p54	11.06
pr94	8.85
p134	6.71
pr174	5.66
p214	5.90
p17	7.92
p57	8.70
p97	5.54
p137	5.66
p19	5.80
p143	5.14
p183	6.52
p223	5.26

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix I.4

New BOR Scale

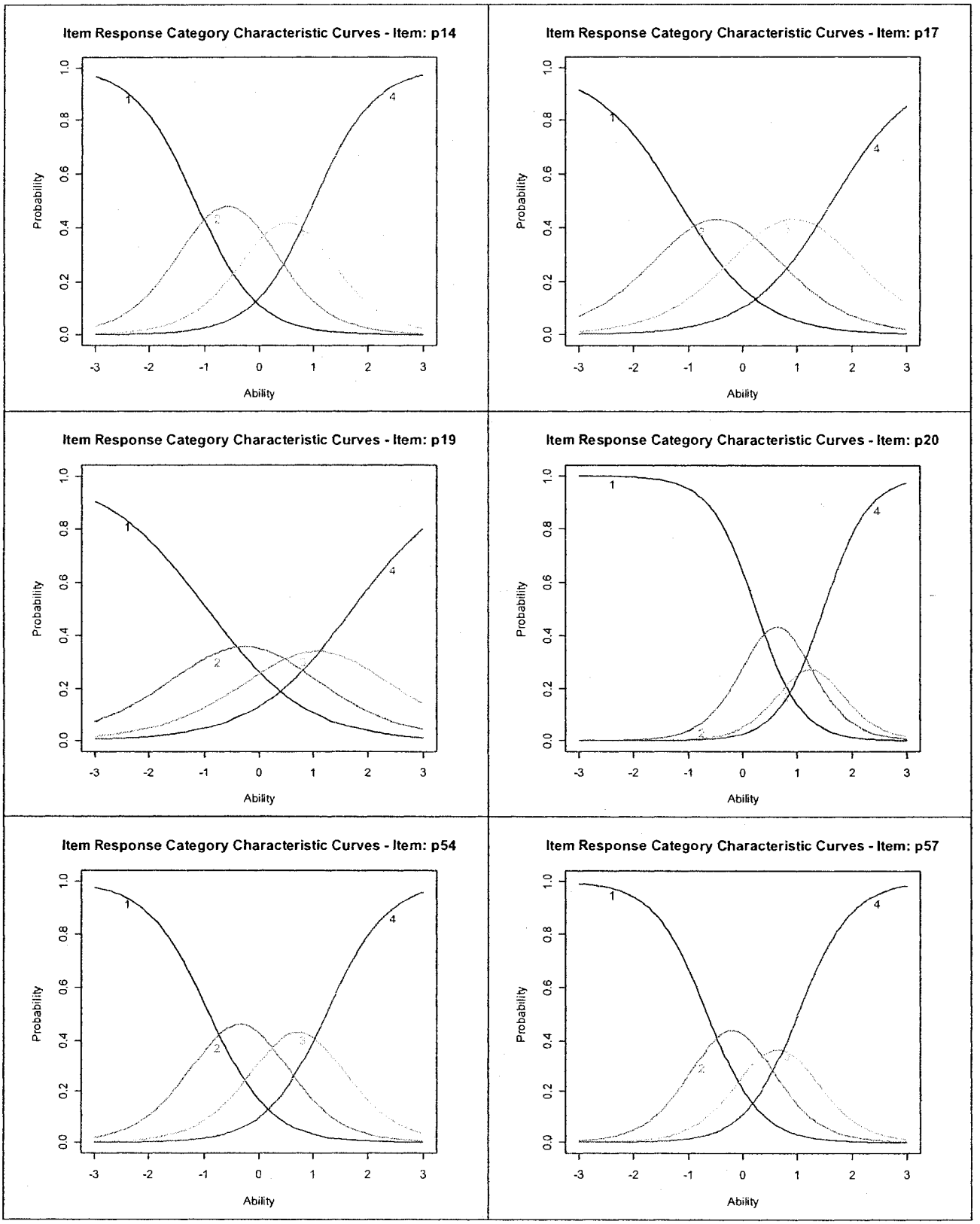

Figure 40. Item Response Category Characteristic Curves (CCC) for BOR-NEW Items (1-6)

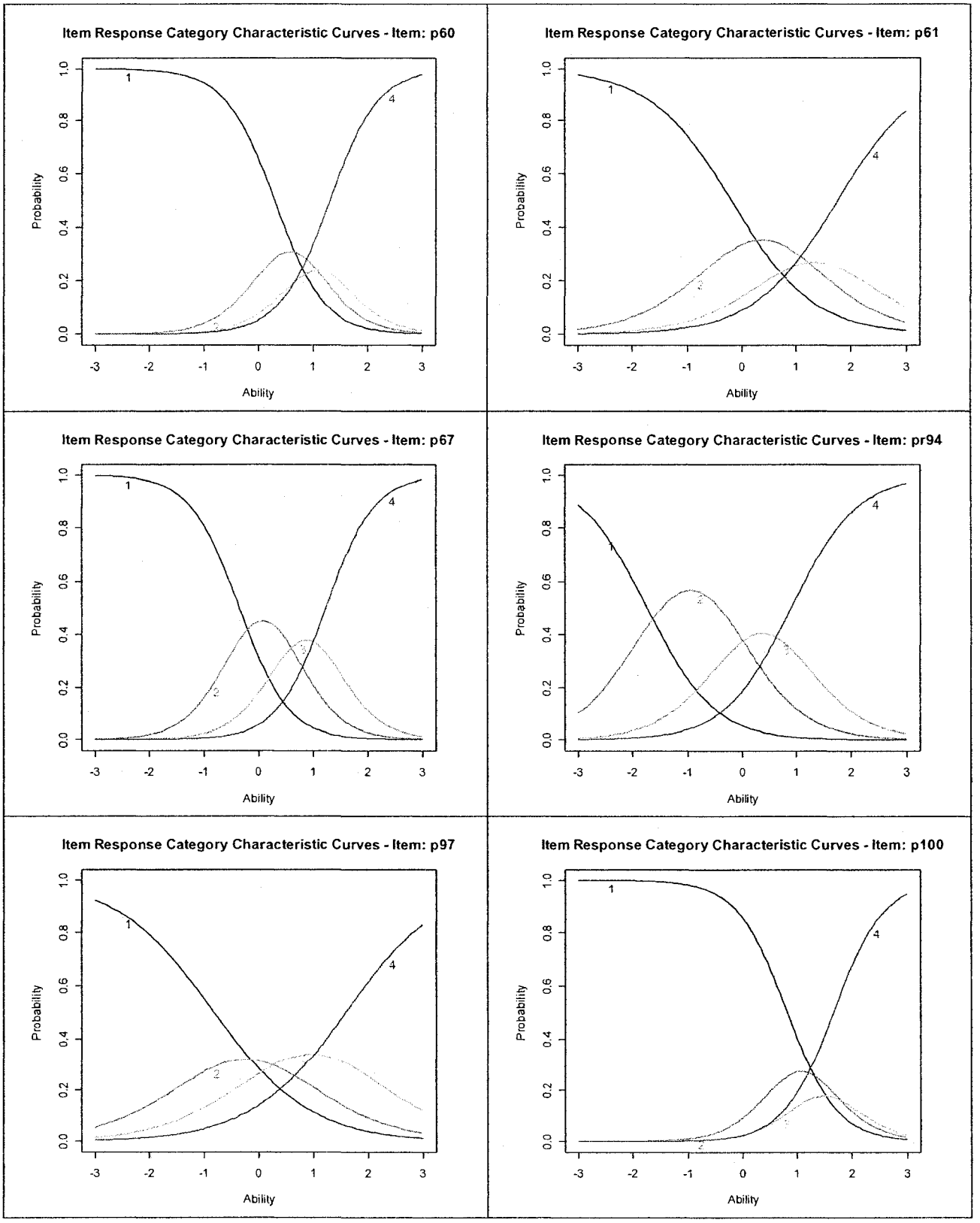

Figure 40 cont'd. Item Response CCCs for BOR-NEW Items (7-12)

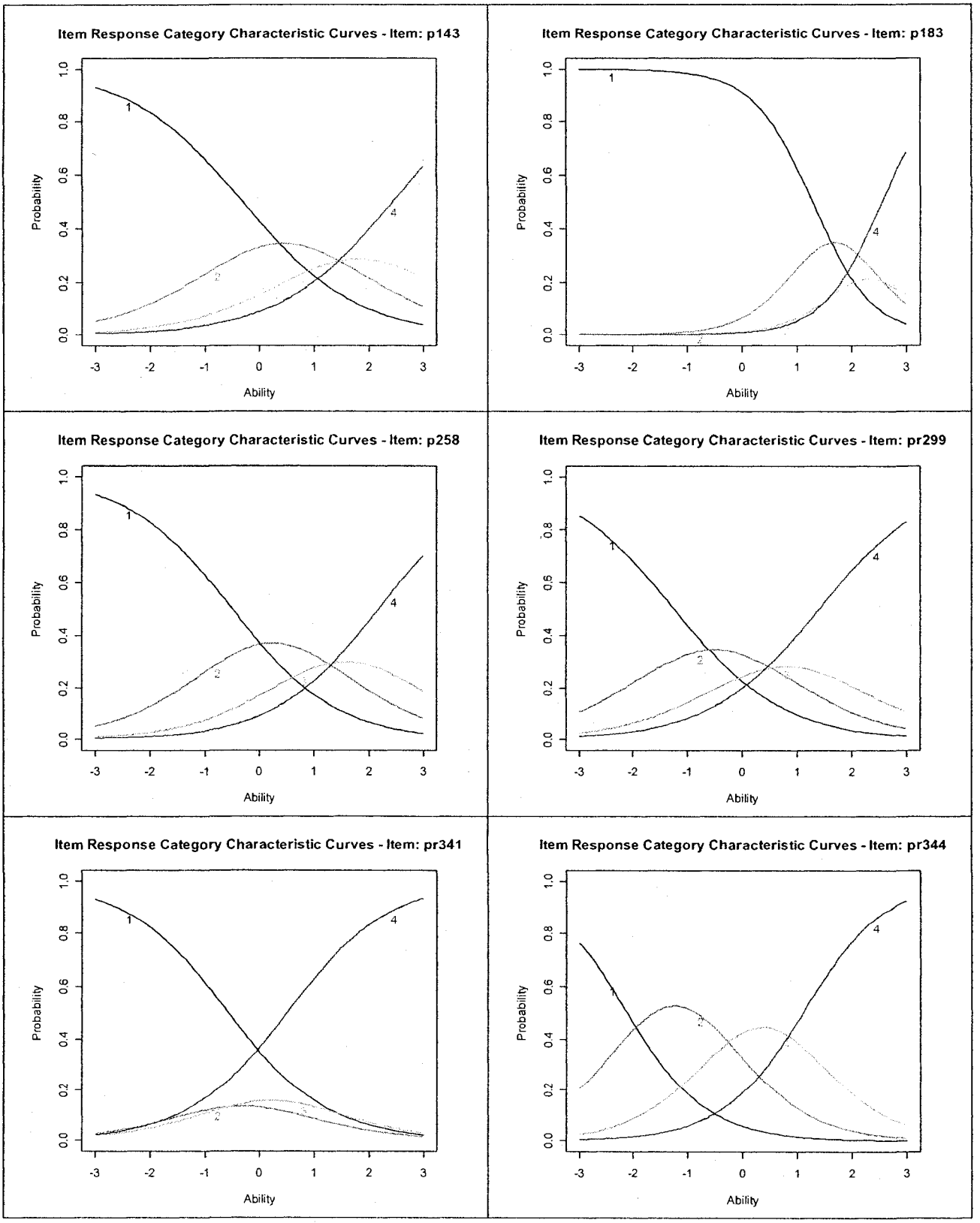

Figure 40 cont'd. Item Response CCCs for BOR-NEW Items (13-18)

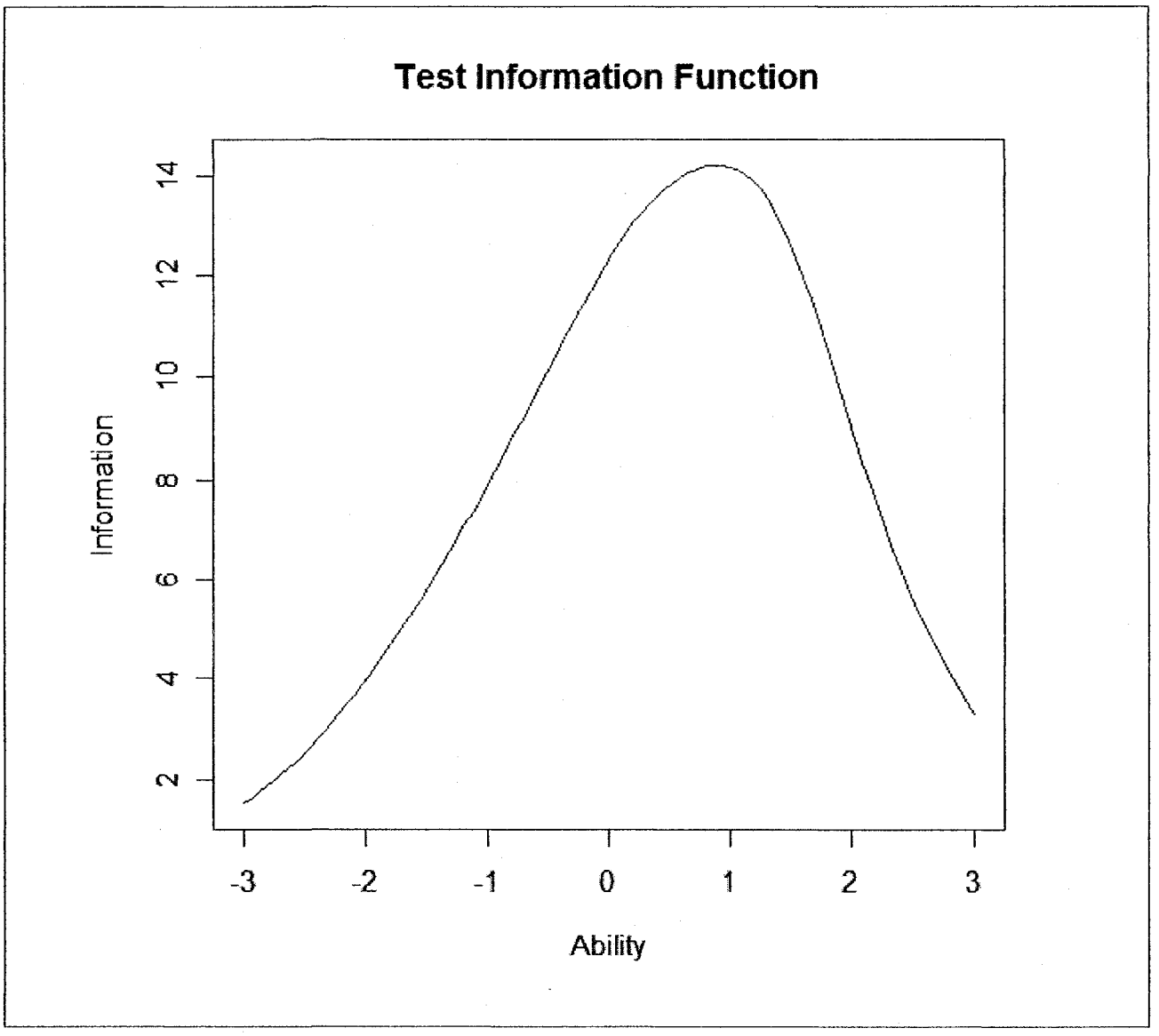

Figure 41. Test Information Function for the new BOR scale.

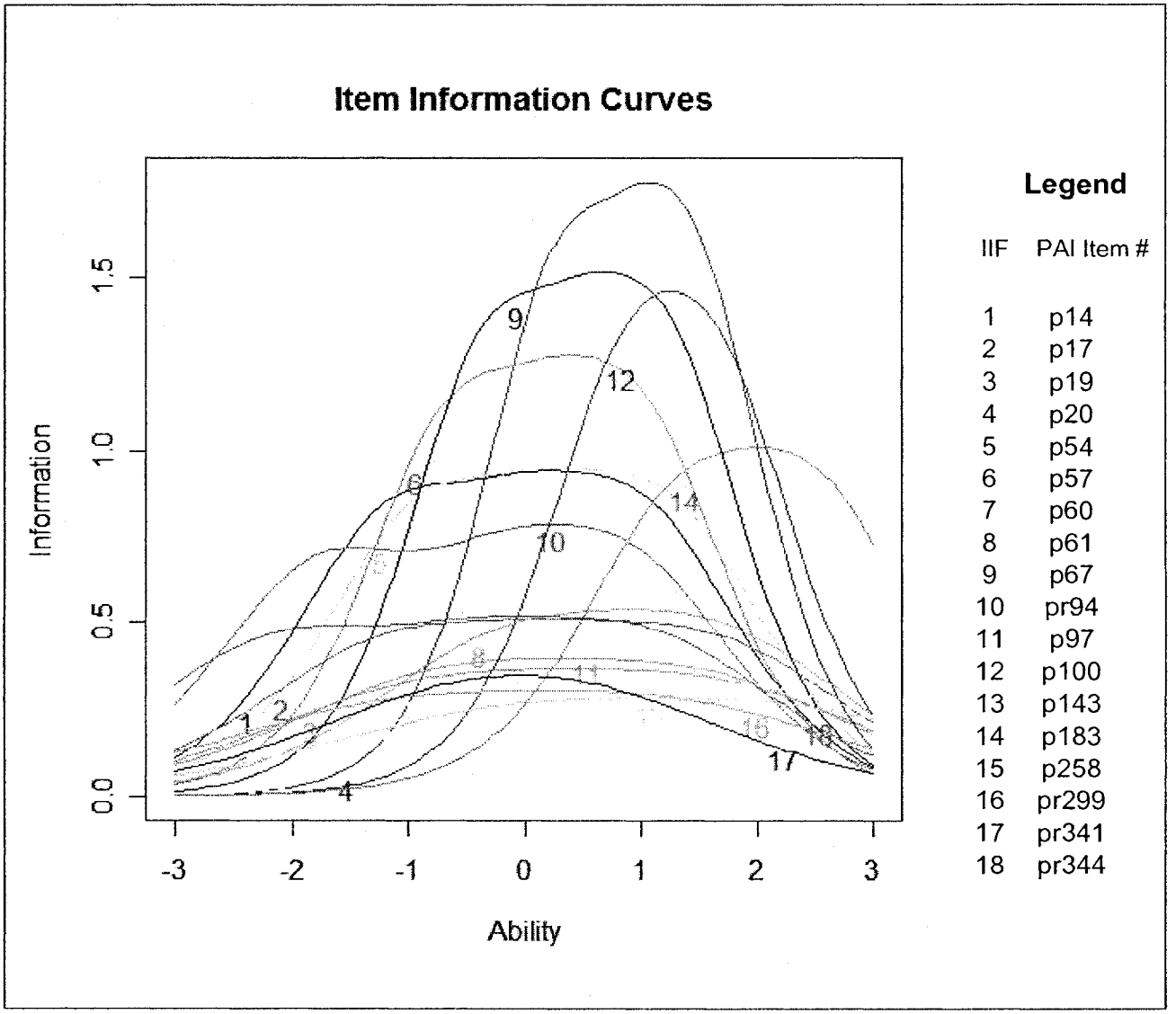

Figure 42. Item Information Functions for the new BOR scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 17.1

New BOR Scale: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p14	-1.16	0.02	1.02	1.78	.46
p17	-1.18	0.23	1.65	1.31	.39
p19	-0.94	0.43	1.73	1.09	.37
p20	0.25	1.02	1.48	2.40	.48
p54	-0.90	0.22	1.25	1.79	.46
p57	-0.65	0.27	1.02	2.05	.47
p60	0.29	0.87	1.31	2.20	.47
p61	-0.20	0.92	1.76	1.32	.43
p67	-0.35	0.53	1.24	2.24	.47
pr94	-1.73	-0.15	0.90	1.63	.42
p97	-0.82	0.35	1.60	1.14	.37
p100	0.83	1.34	1.67	2.17	.47
p143	-0.30	1.20	2.44	0.96	.37
p183	1.29	2.09	2.57	1.81	.51
p258	-0.49	1.00	2.18	1.05	.39
pr299	-1.23	0.23	1.40	0.99	.37
pr341	-0.57	-0.07	0.51	1.08	.35
pr344	-2.11	-0.35	1.09	1.33	.40
Mean	-0.55	0.56	1.49	1.57	*(.42)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 17.2

New BOR Scale: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	92.28
-2 to +2	76.84
-3 to 0	34.08
0 to +3	58.19
-3 to -2	4.78
-2 to -1	10.69
-1 to 0	18.61
0 to +1	25.02
+1 to +2	22.51
+2 to +3	10.66

Note. [1] = BOR trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 17.3

New BOR Scale: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p14	7.04
p17	5.07
p19	3.84
p20	8.45
p54	7.04
p57	7.68
p60	6.95
p61	4.39
p67	8.51
pr94	6.67
p97	3.88
p100	6.36
p143	3.20
p183	5.77
p258	3.60
pr299	3.33
pr341	2.74
pr344	5.48

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix J

HIS

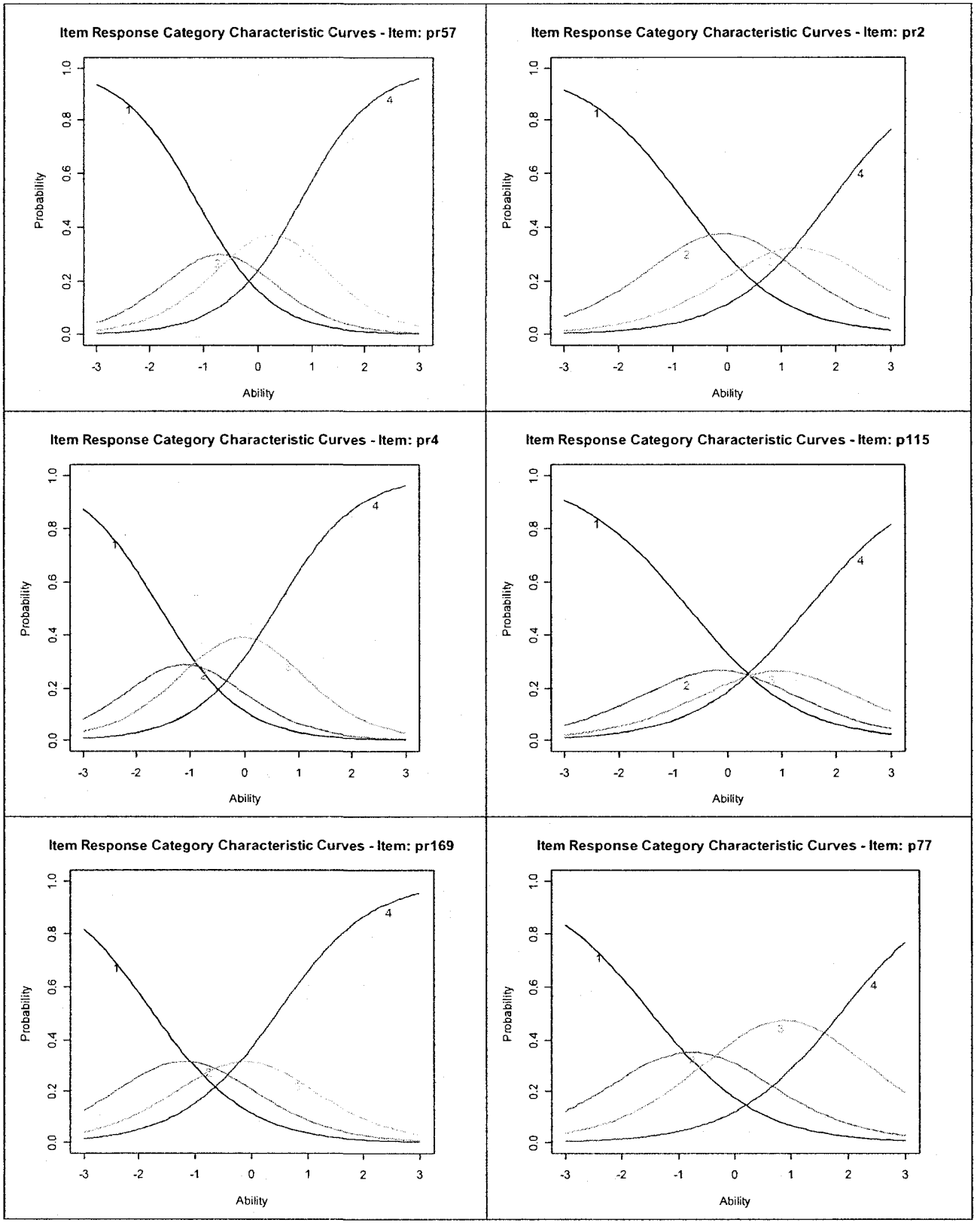

Figure 43. Item Response Category Characteristic Curves (CCC) for HIS Items (1-6)

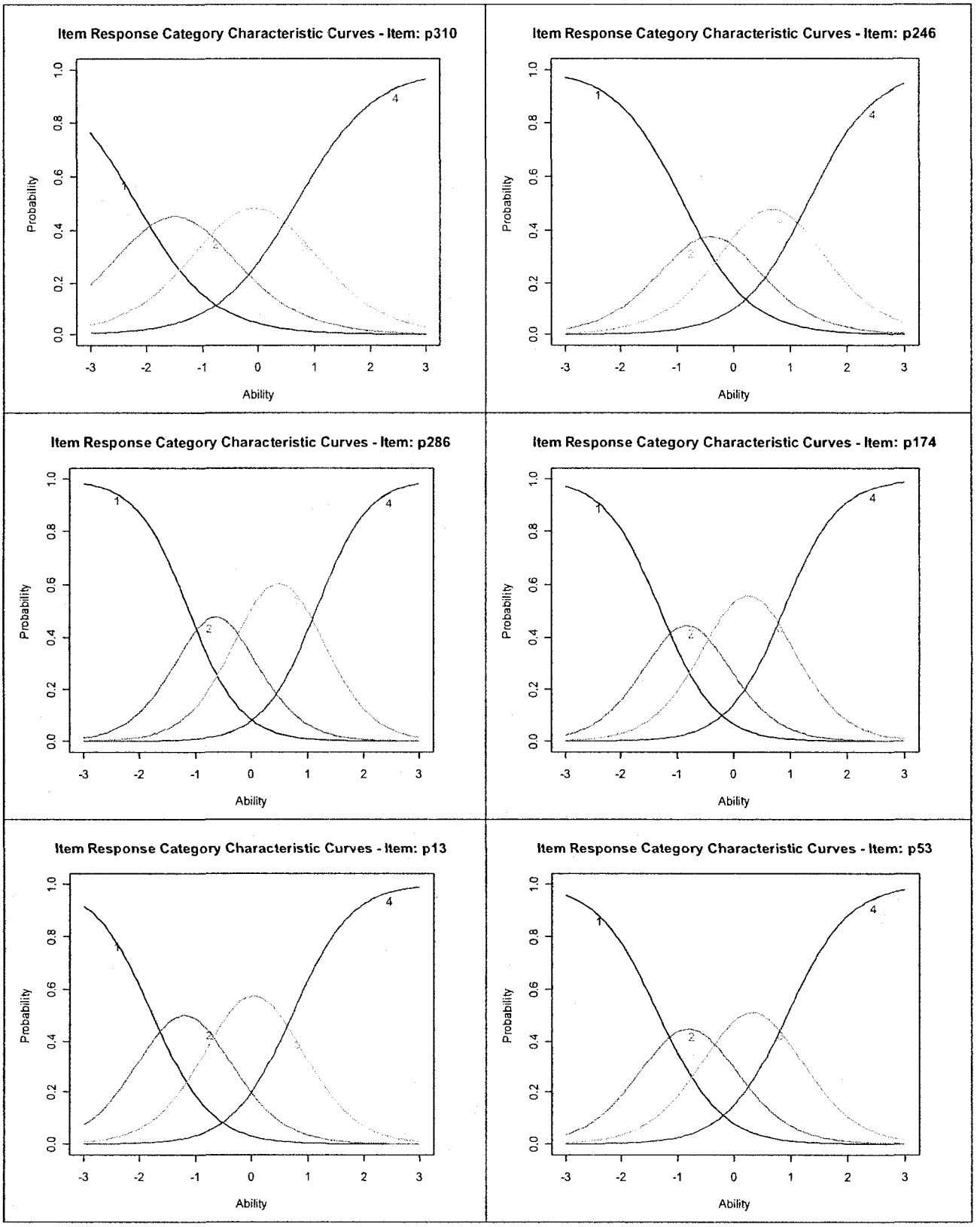

Figure 43 cont'd. Item Response Category Characteristic Curves (CCC) for HIS Items (7-12)

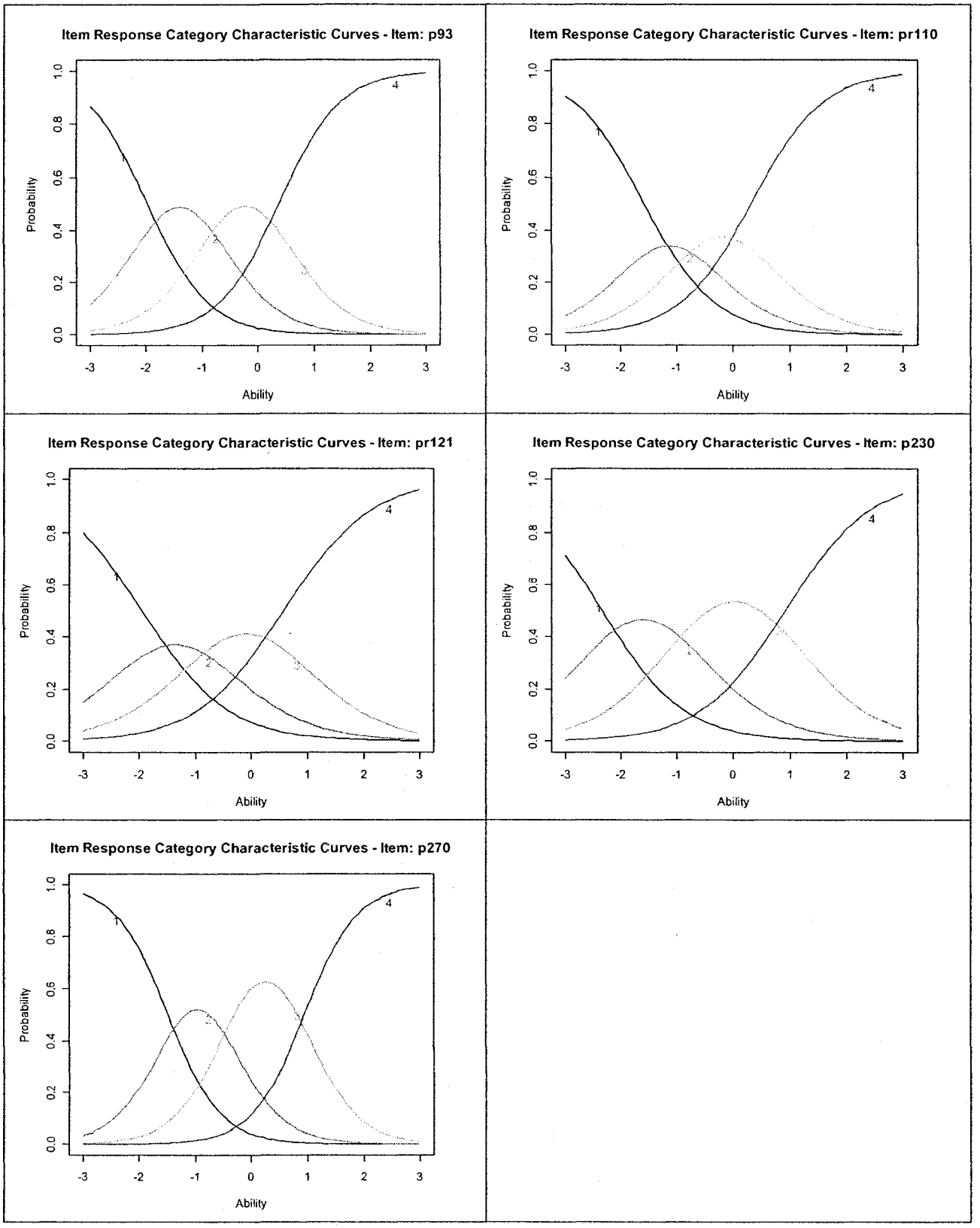

Figure 43 cont'd. Item Response Category Characteristic Curves (CCC) for HIS Items (13-17)

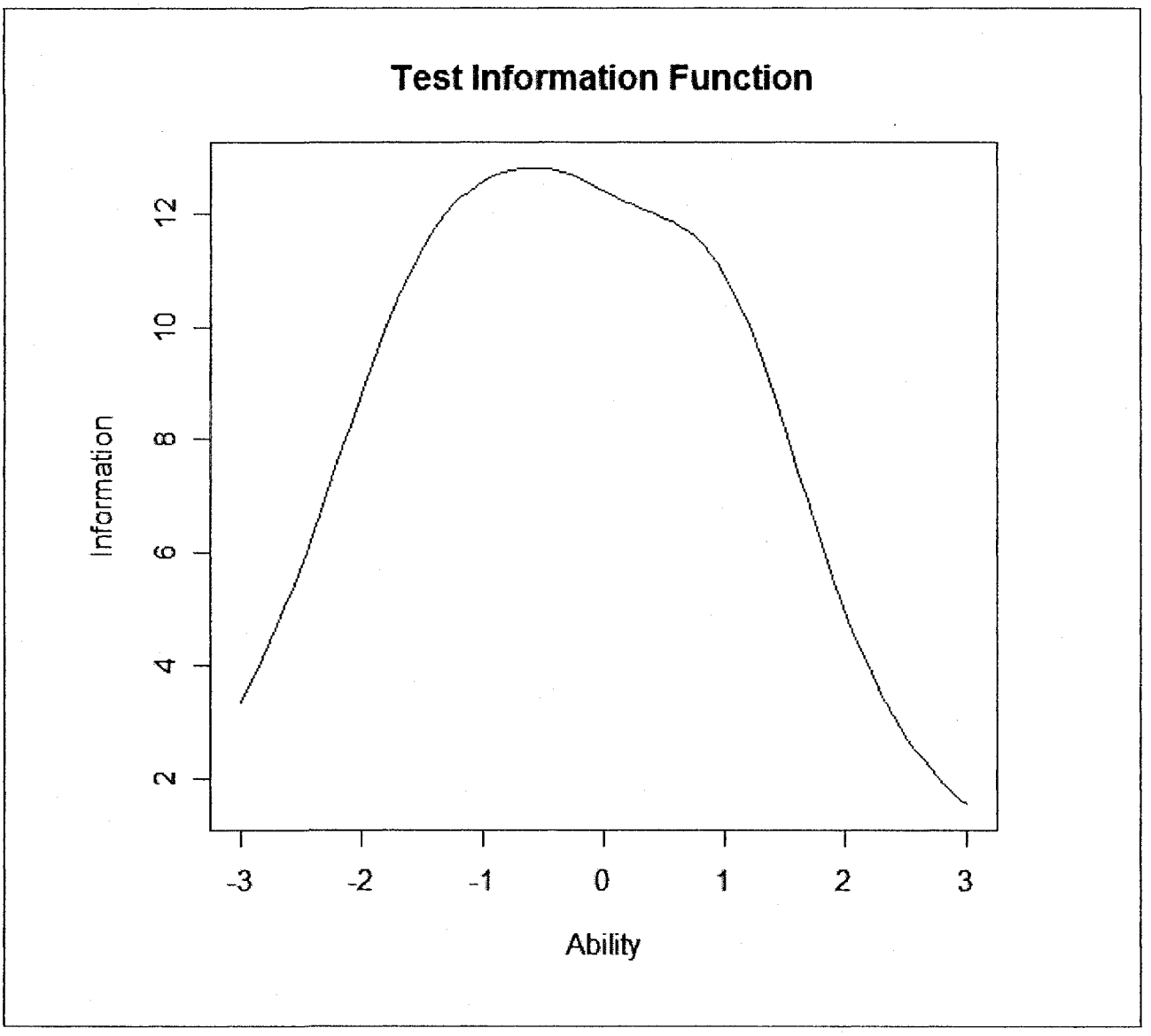

Figure 44. Test Information Function for the HIS scale.

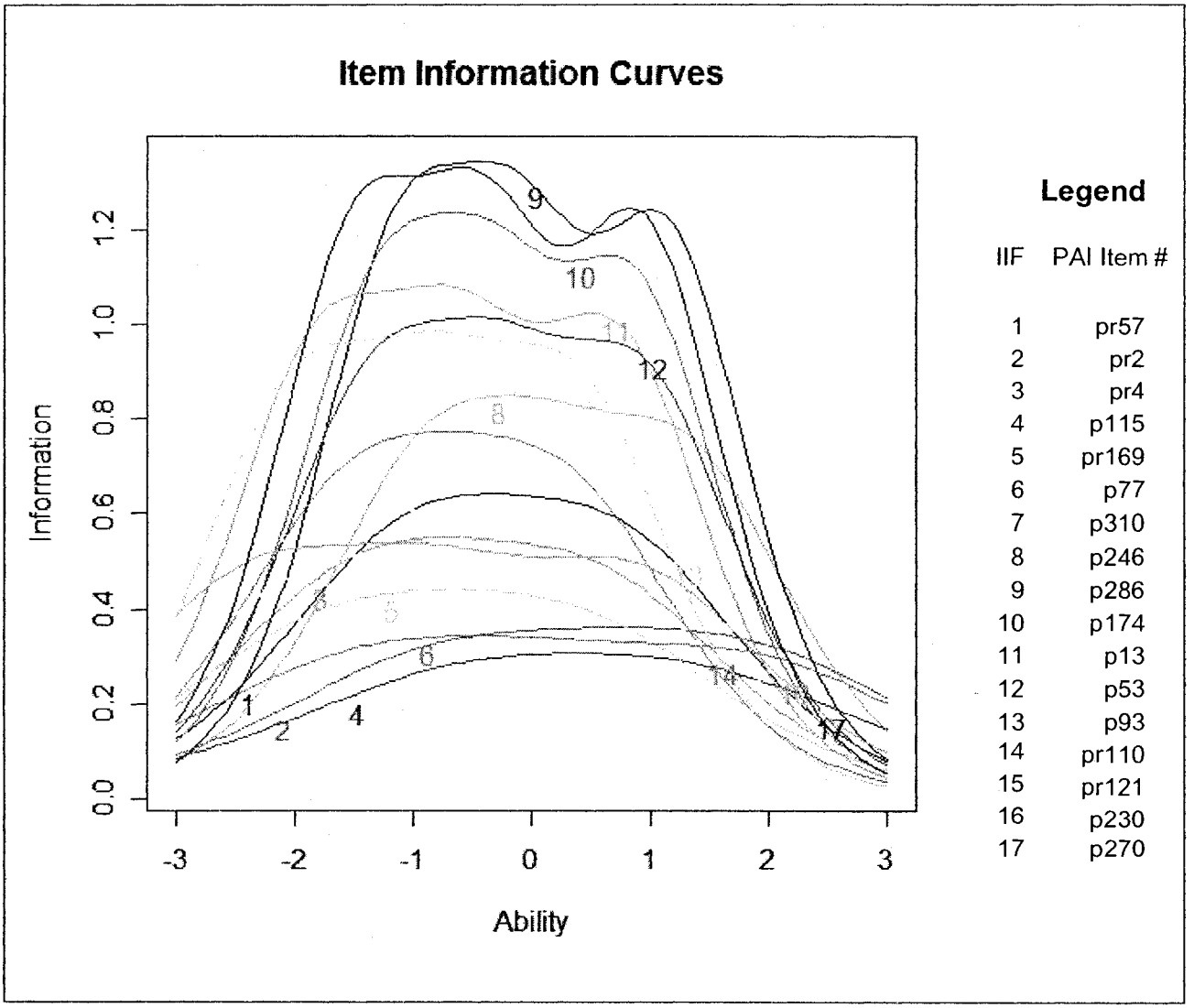

Figure 45. Item Information Functions for the HIS scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 18.1

HIS: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
pr57	-1.14	-0.29	0.80	1.44	.45
pr2	-0.81	0.66	1.91	1.08	.38
pr4	-1.55	-0.66	0.58	1.34	.42
p115	-0.71	0.39	1.48	0.99	.34
pr169	-1.73	-0.64	0.44	1.19	.39
p77	-1.47	-0.08	1.86	1.06	.38
p310	-2.18	-0.81	0.66	1.43	.44
p246	-0.89	0.06	1.30	1.68	.46
p286	-1.12	-0.16	1.14	2.16	.51
p174	-1.29	-0.36	0.87	2.05	.50
p13	-1.76	-0.63	0.72	1.95	.44
p53	-1.33	-0.29	0.93	1.86	.44
p93	-1.97	-0.81	0.37	1.84	.44
pr110	-1.57	-0.68	0.32	1.58	.43
pr121	-1.94	-0.76	0.57	1.31	.40
p230	-2.34	-0.85	0.91	1.36	.39
p270	-1.48	-0.42	0.94	2.17	.46
Mean	-1.49	-0.37	0.93	1.56	*(.43)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 18.2

HIS: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	92.96
-2 to +2	77.42
-3 to 0	52.53
0 to +3	40.42
-3 to -2	10.39
-2 to -1	19.69
-1 to 0	22.46
0 to +1	20.98
+1 to +2	14.29
+2 to +3	5.15

Note. [1] = HIS trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 18.3

HIS: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
pr57	4.76
pr2	3.66
pr4	4.44
p115	2.97
pr169	3.82
p77	3.89
p310	5.55
p246	6.23
p286	8.95
p174	8.21
p13	8.08
p53	7.31
p93	7.31
pr110	5.40
pr121	4.65
p230	5.47
p270	9.30

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix K

NAR

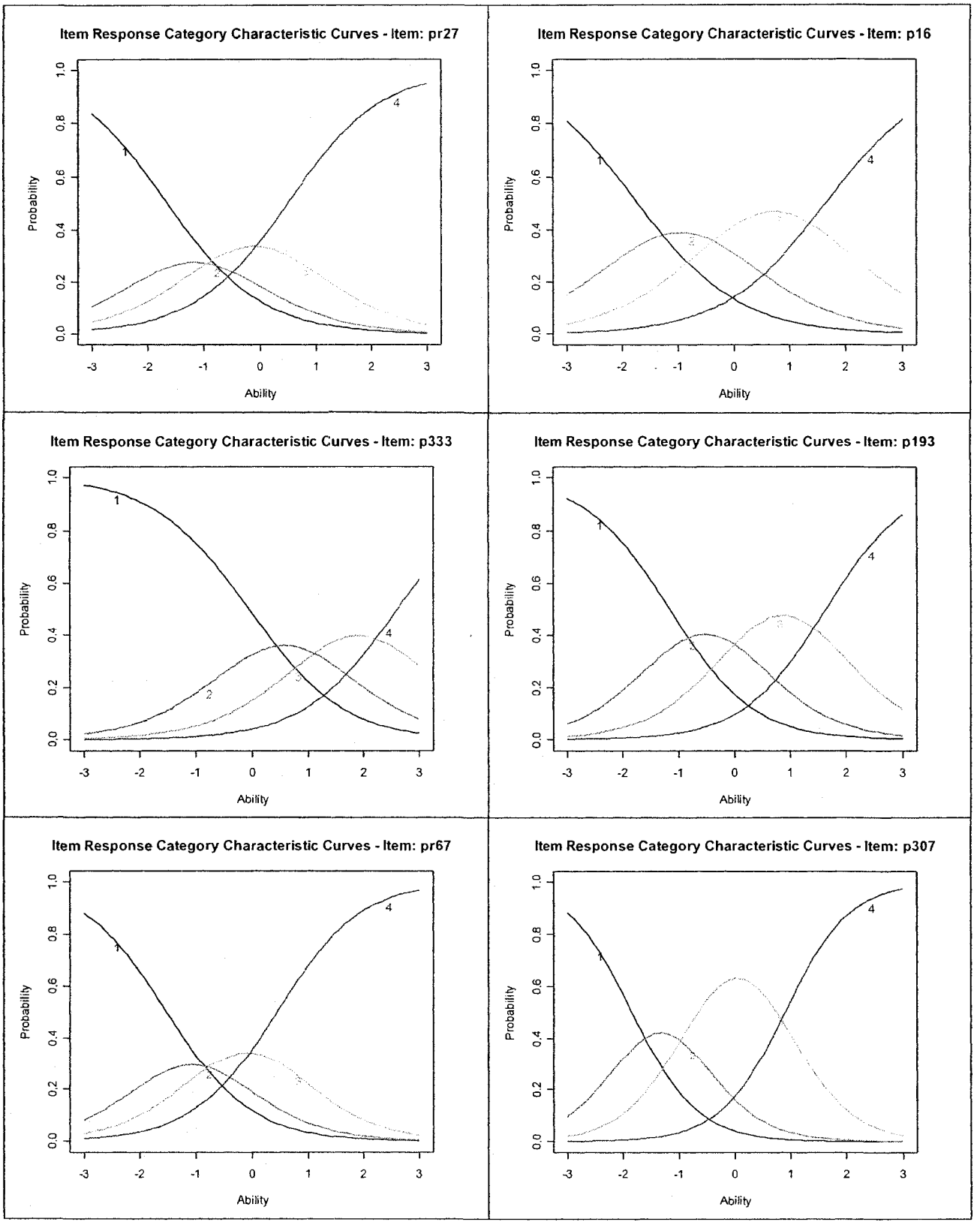

Figure 46. Item Response Category Characteristic Curves (CCC) for NAR Items (1-6)

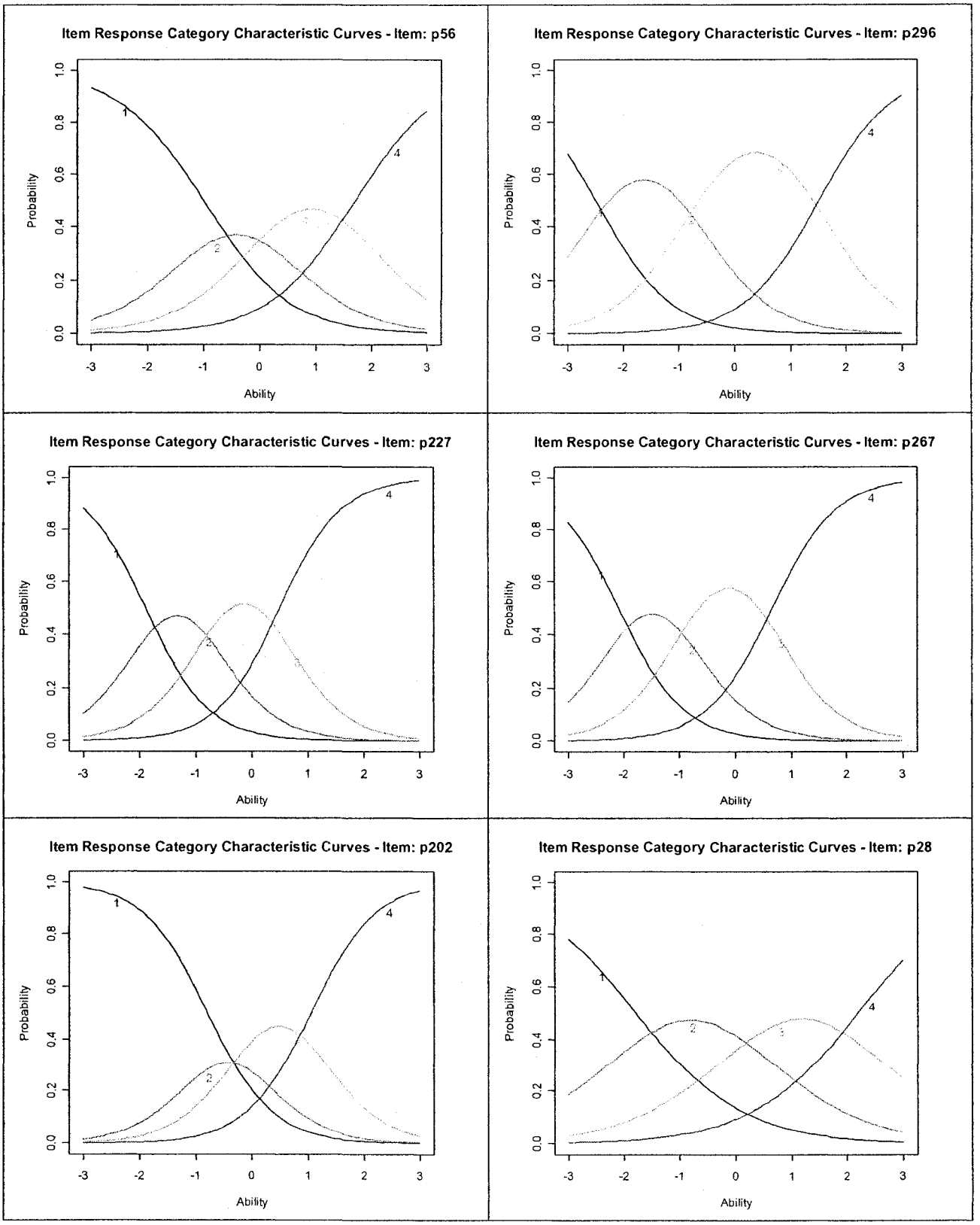

Figure 46 cont'd. Item Response Category Characteristic Curves (CCC) for NAR Items (7-12)

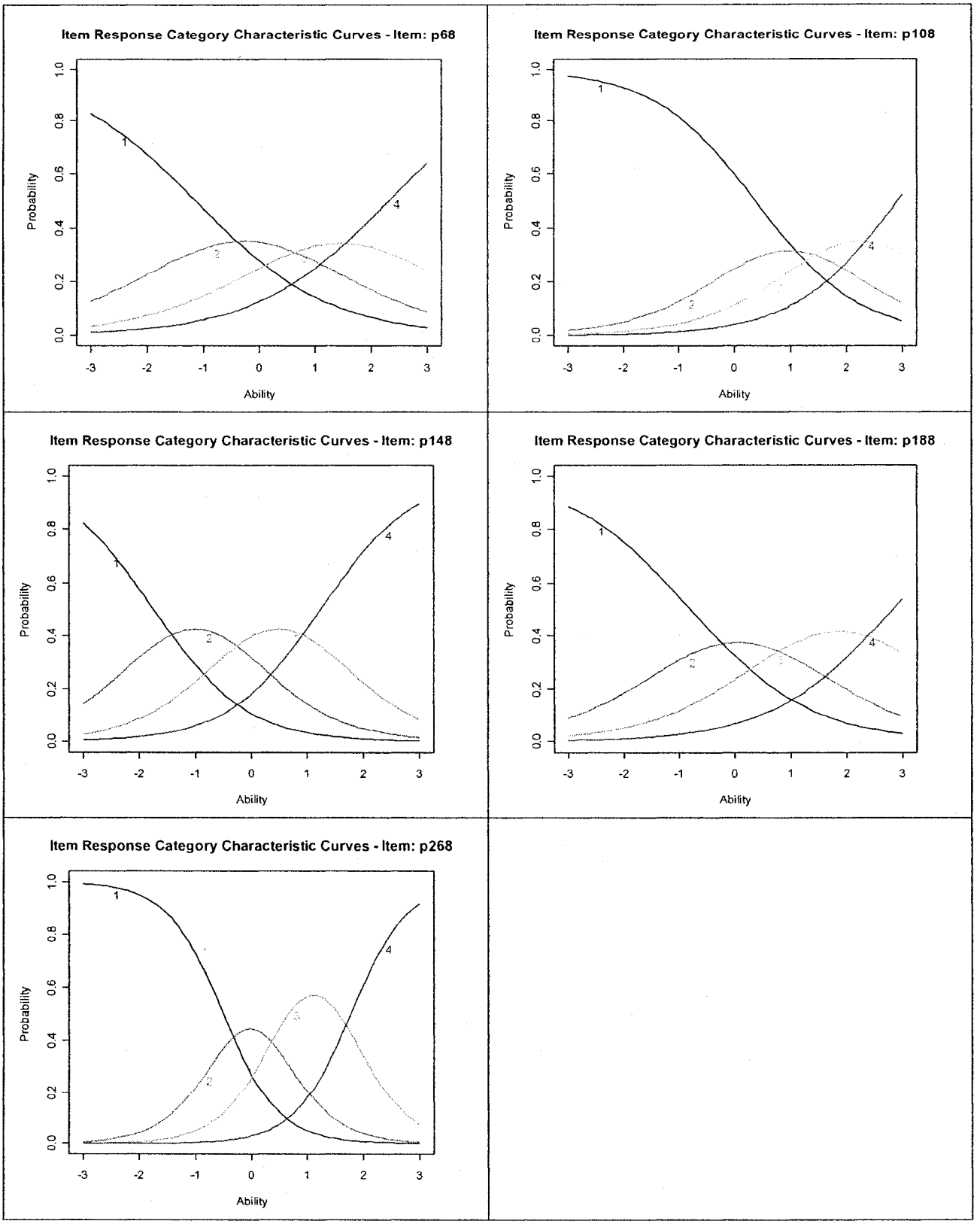

Figure 46 cont'd. Item Response Category Characteristic Curves (CCC) for NAR Items (13-17)

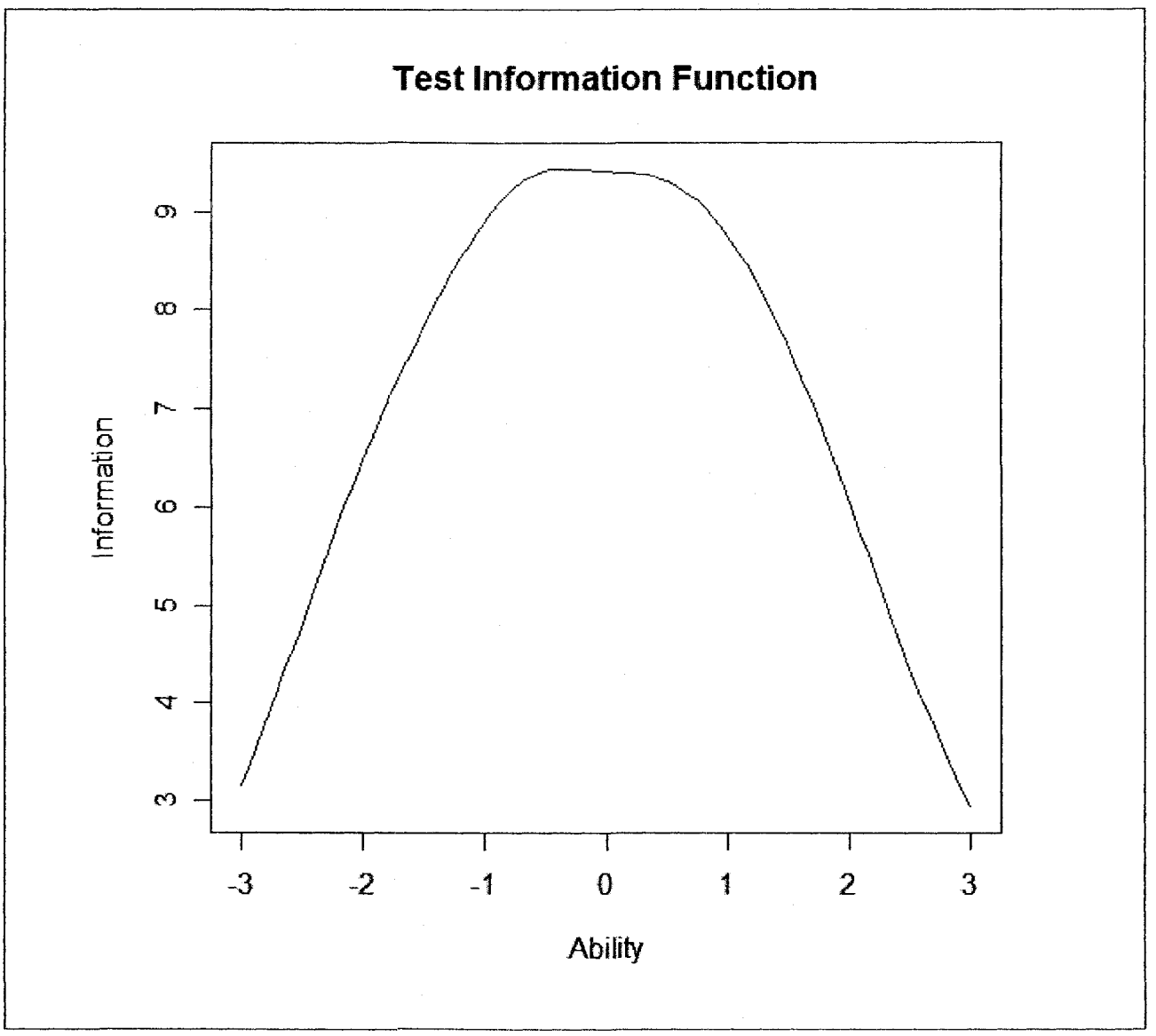

Figure 47. Test Information Function for the NAR scale.

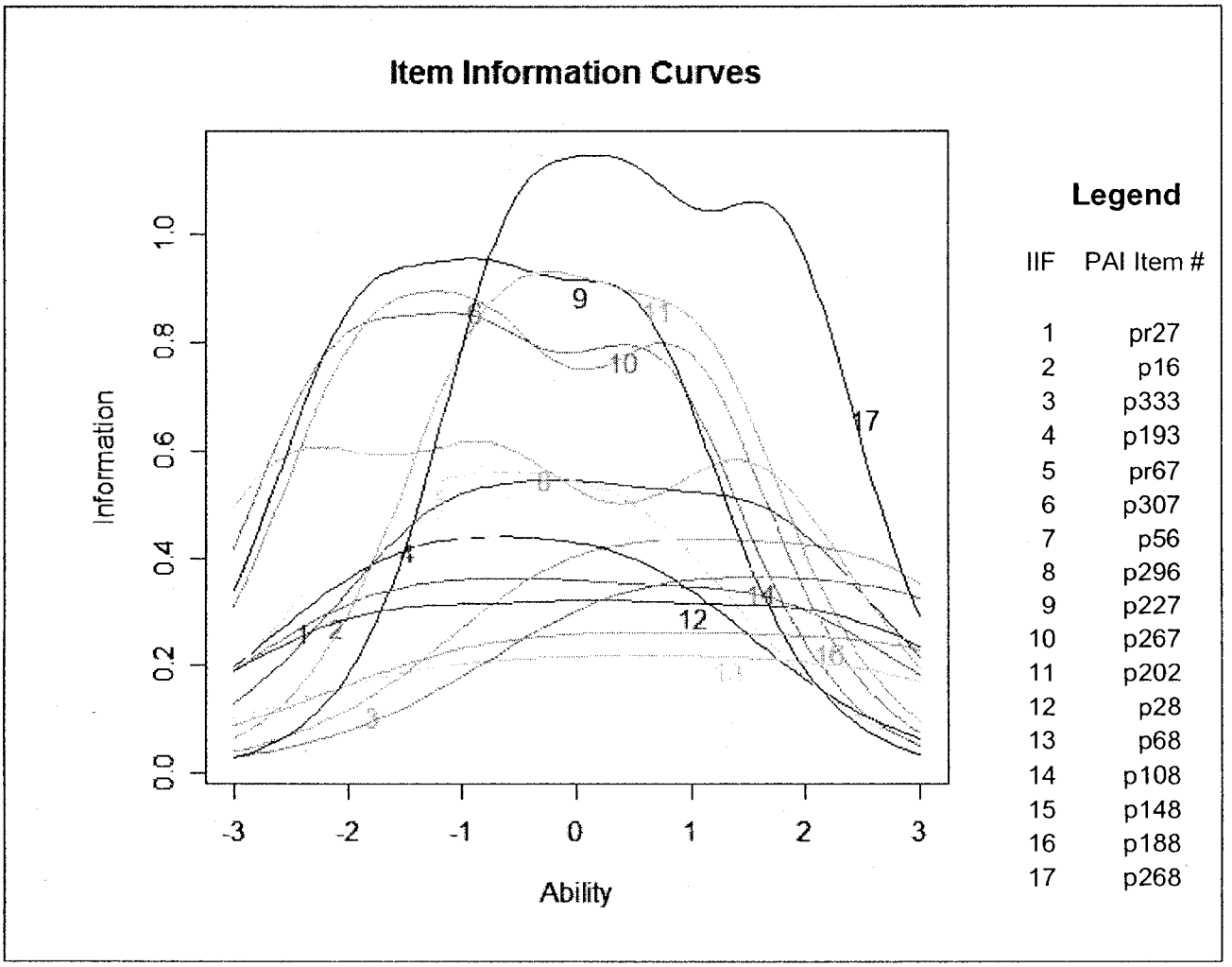

Figure 48. Item Information Functions for the NAR scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 19.1

NAR: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
pr27	-1.63	-0.68	0.49	1.19	.35
p16	-1.71	-0.21	1.64	1.10	.35
p333	-0.06	1.20	2.61	1.20	.37
p193	-1.15	0.11	1.64	1.35	.39
pr67	-1.51	-0.62	0.44	1.35	.37
p307	-1.83	-0.80	0.90	1.74	.43
p56	-1.01	0.17	1.70	1.32	.38
p296	-2.49	-0.73	1.50	1.50	.40
p227	-1.89	-0.76	0.50	1.81	.41
p267	-2.08	-0.87	0.66	1.72	.43
p202	-0.79	-0.07	1.06	1.75	.42
p28	-1.78	0.19	2.19	1.05	.35
p68	-1.13	0.61	2.31	0.84	.32
p108	0.37	1.57	2.91	1.09	.38
p148	-1.74	-0.25	1.24	1.22	.37
p188	-0.80	0.90	2.81	0.93	.33
p268	-0.50	0.46	1.77	1.98	.45
Mean	-1.28	0.01	1.55	1.36	*(.38)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 19.2

NAR: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	87.65
-2 to +2	68.89
-3 to 0	44.57
0 to +3	43.07
-3 to -2	9.80
-2 to -1	15.79
-1 to 0	18.98
0 to +1	18.79
+1 to +2	15.34
+2 to +3	8.95

Note. [1] = NAR trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 19.3

NAR: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
pr27	4.38
p16	4.70
p333	4.82
p193	5.84
pr67	5.01
p307	8.20
p56	5.56
p296	7.78
p227	8.26
p267	8.12
p202	7.04
p28	4.72
p68	3.26
p108	4.13
p148	5.21
p188	3.81
p268	9.16

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix L

DEP

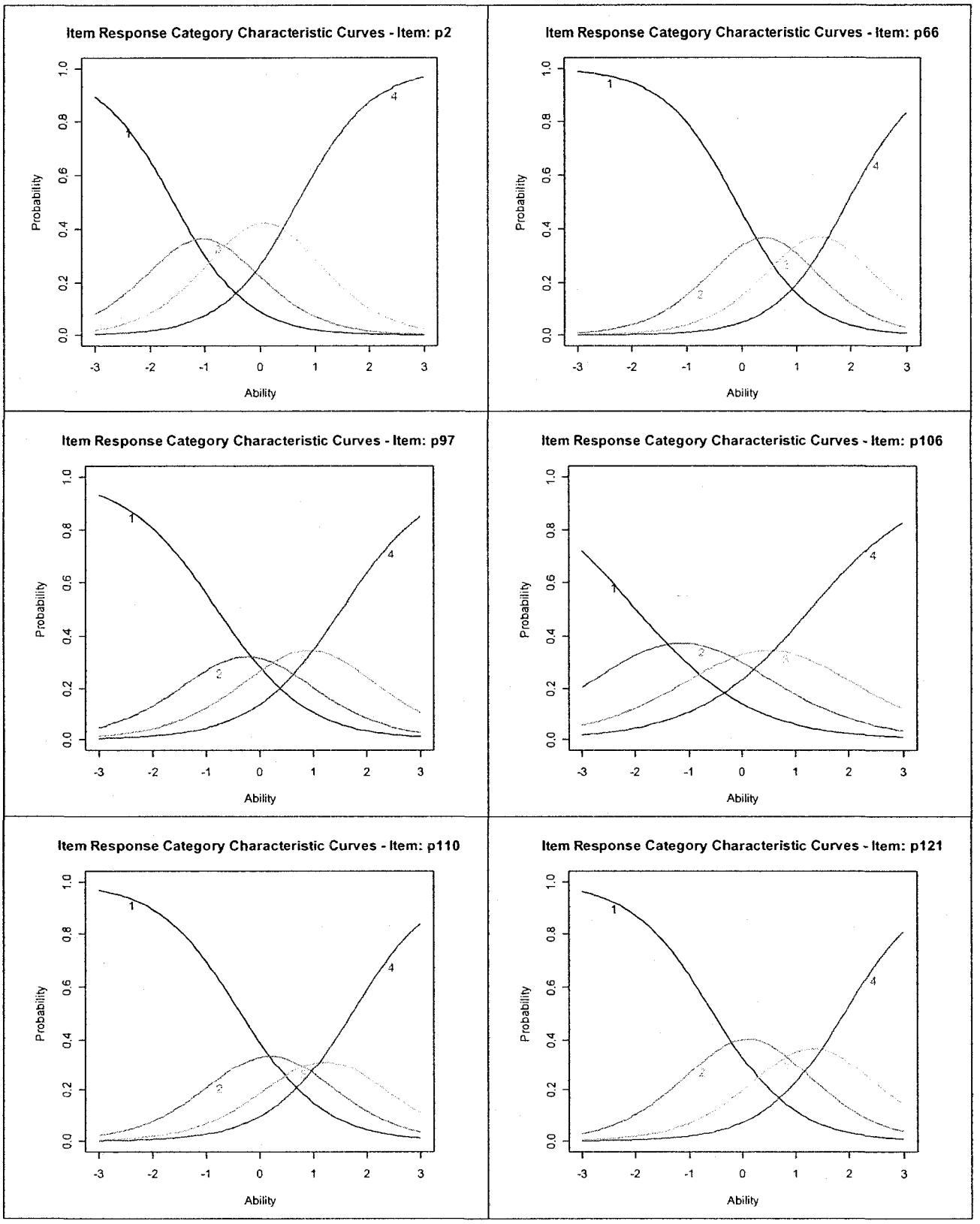

Figure 49. Item Response Category Characteristic Curves (CCC) for DEP Items (1-6)

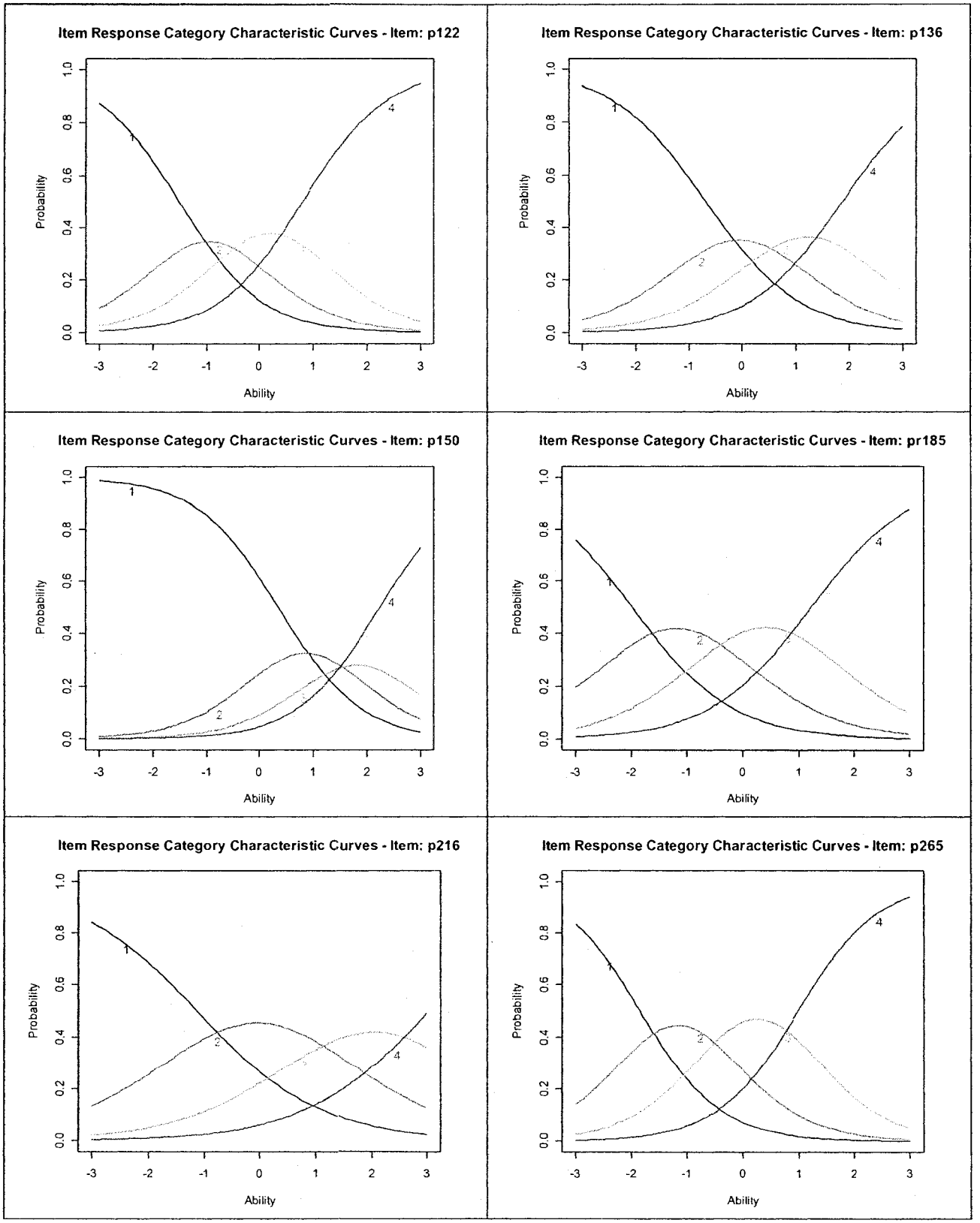

Figure 49 cont'd. Item Response Category Characteristic Curves (CCC) for DEP Items (7-12)

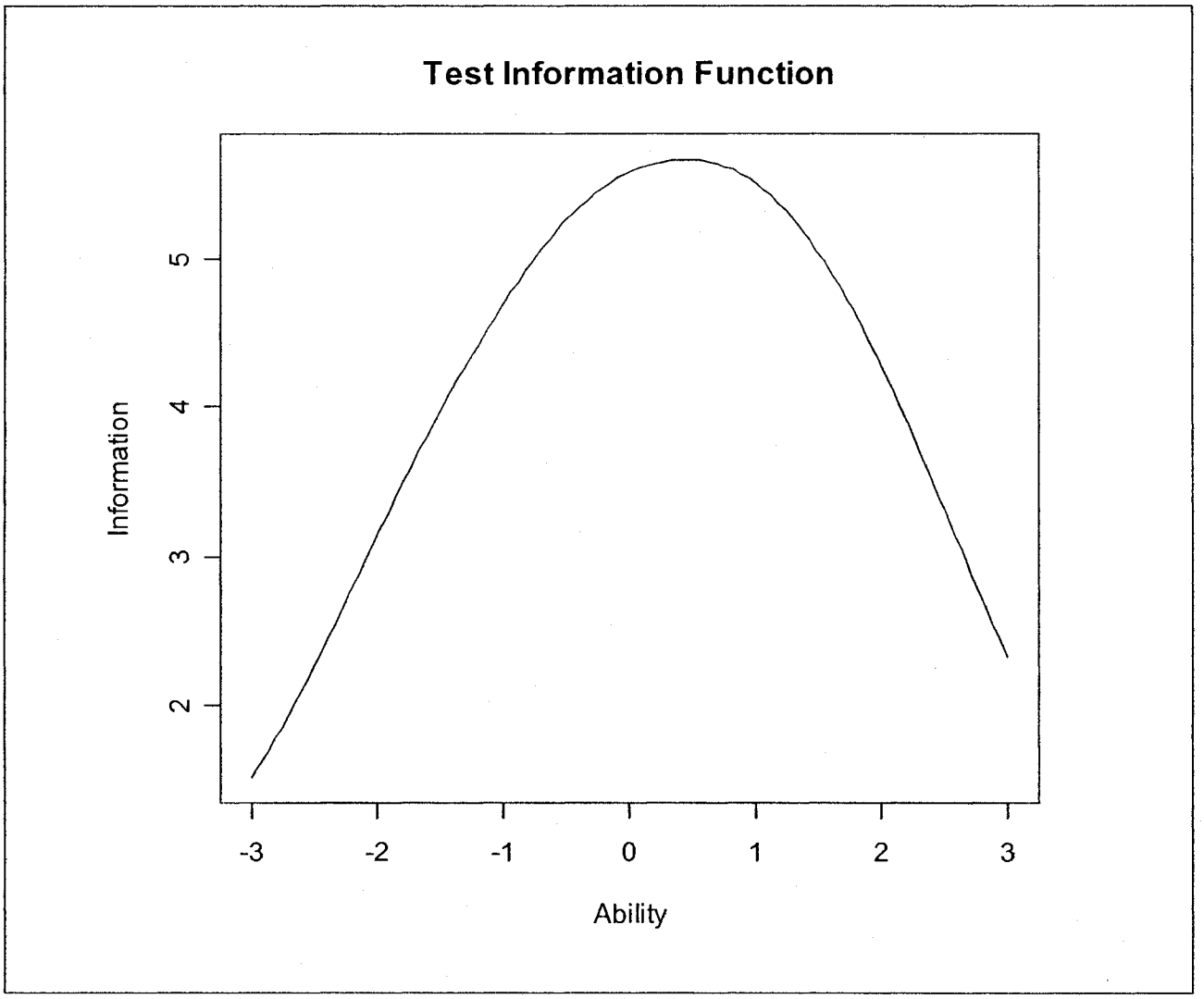

Figure 50. Test Information Function for the DEP scale.

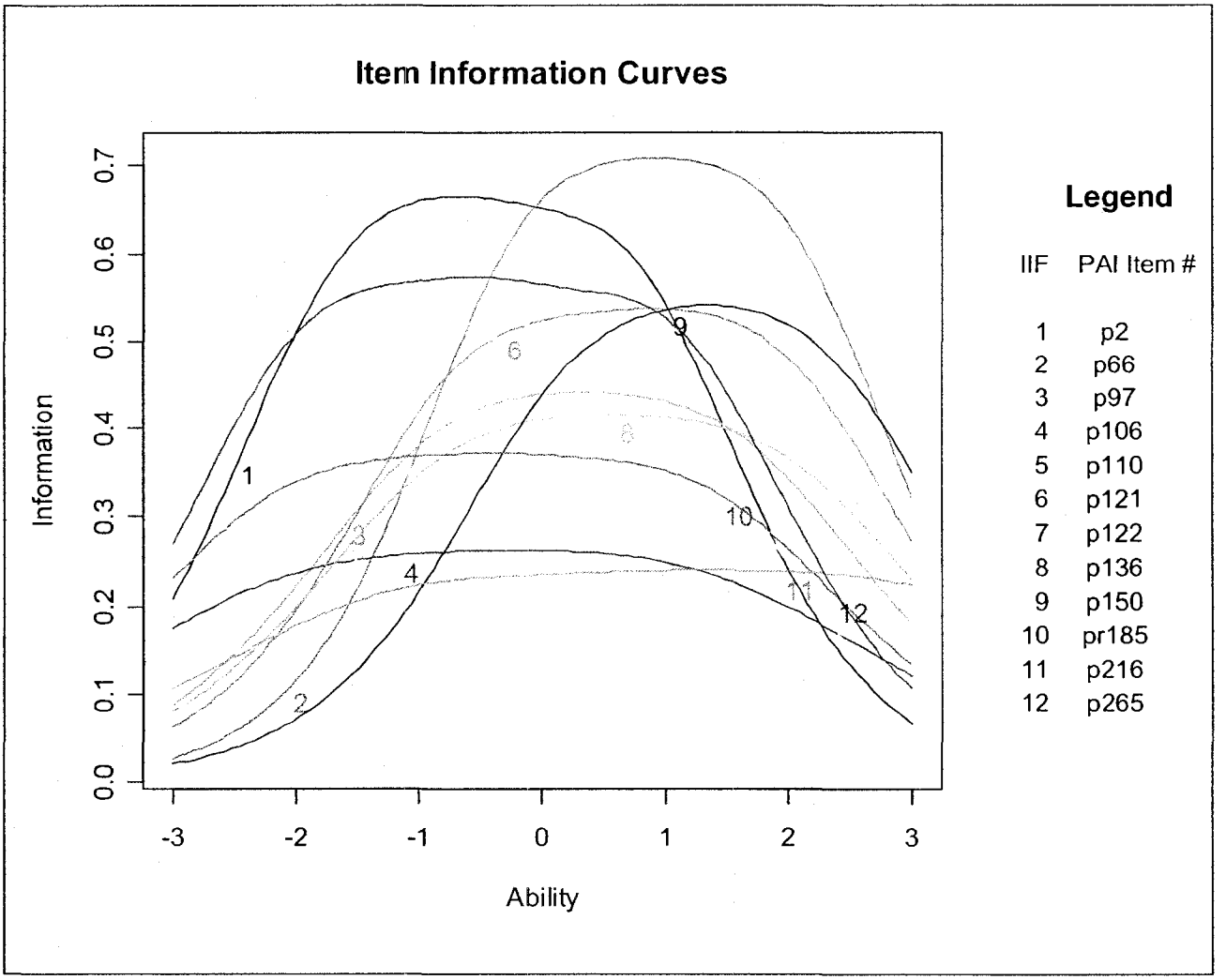

Figure 51. Item Information Functions for the DEP scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 20.1

DEP: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
p2	-1.56	-0.53	0.69	1.48	.35
p66	-0.10	0.92	1.94	1.52	.35
p97	-0.79	0.32	1.53	1.19	.31
p106	-1.97	-0.27	1.28	0.92	.26
p110	-0.35	0.72	1.72	1.30	.32
p121	-0.54	0.75	1.91	1.33	.33
p122	-1.51	-0.40	0.81	1.30	.32
p136	-0.68	0.57	1.88	1.16	.31
p150	0.35	1.37	2.24	1.32	.34
pr185	-1.99	-0.38	1.23	1.11	.28
p216	-1.13	1.05	3.04	0.90	.26
p265	-1.84	-0.47	1.00	1.39	.34
Mean	-1.01	0.31	1.61	1.24	*(.31)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 20.2

DEP: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	86.51
-2 to +2	67.43
-3 to 0	39.04
0 to +3	47.47
-3 to -2	7.80
-2 to -1	13.45
-1 to 0	17.79
0 to +1	19.16
+1 to +2	17.02
+2 to +3	11.29

Note. [1] = DEP trait range in *SD* units, $M = 0$, $SD = 1$; Percent = percent of total information or total area under the Test Information Function.

Table 20.3

DEP: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
p2	10.15
p66	10.12
p97	7.60
p106	6.06
p110	8.21
p121	9.03
p122	8.59
p136	7.60
p150	8.04
pr185	7.87
p216	6.47
p265	10.29

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; Percent = percent of total information or total area under the Item Information Function.

Appendix M

COM

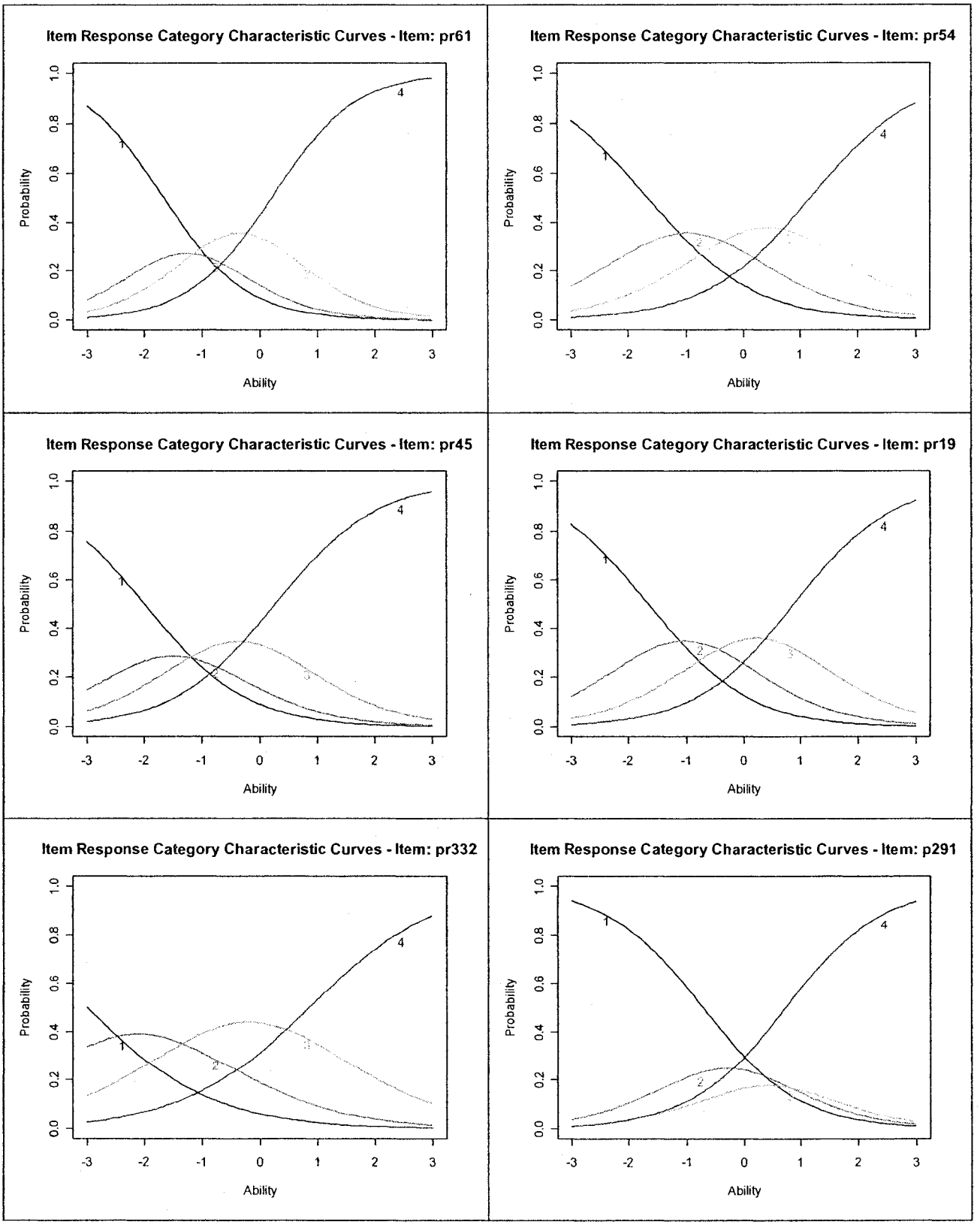

Figure 52. Item Response Category Characteristic Curves (CCC) for COM Items (1-6)

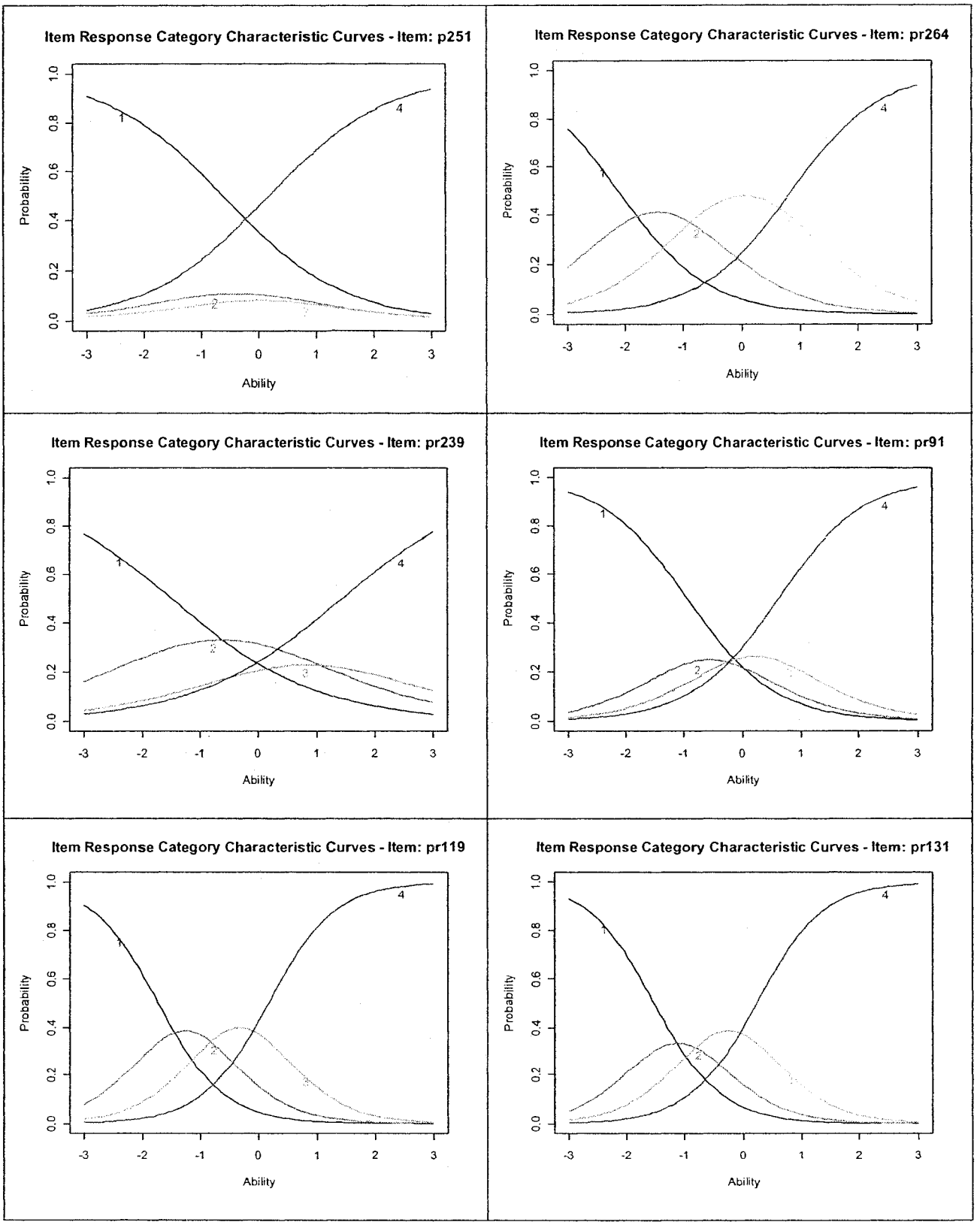

Figure 52 cont'd. Item Response Category Characteristic Curves (CCC) for COM Items (7-12)

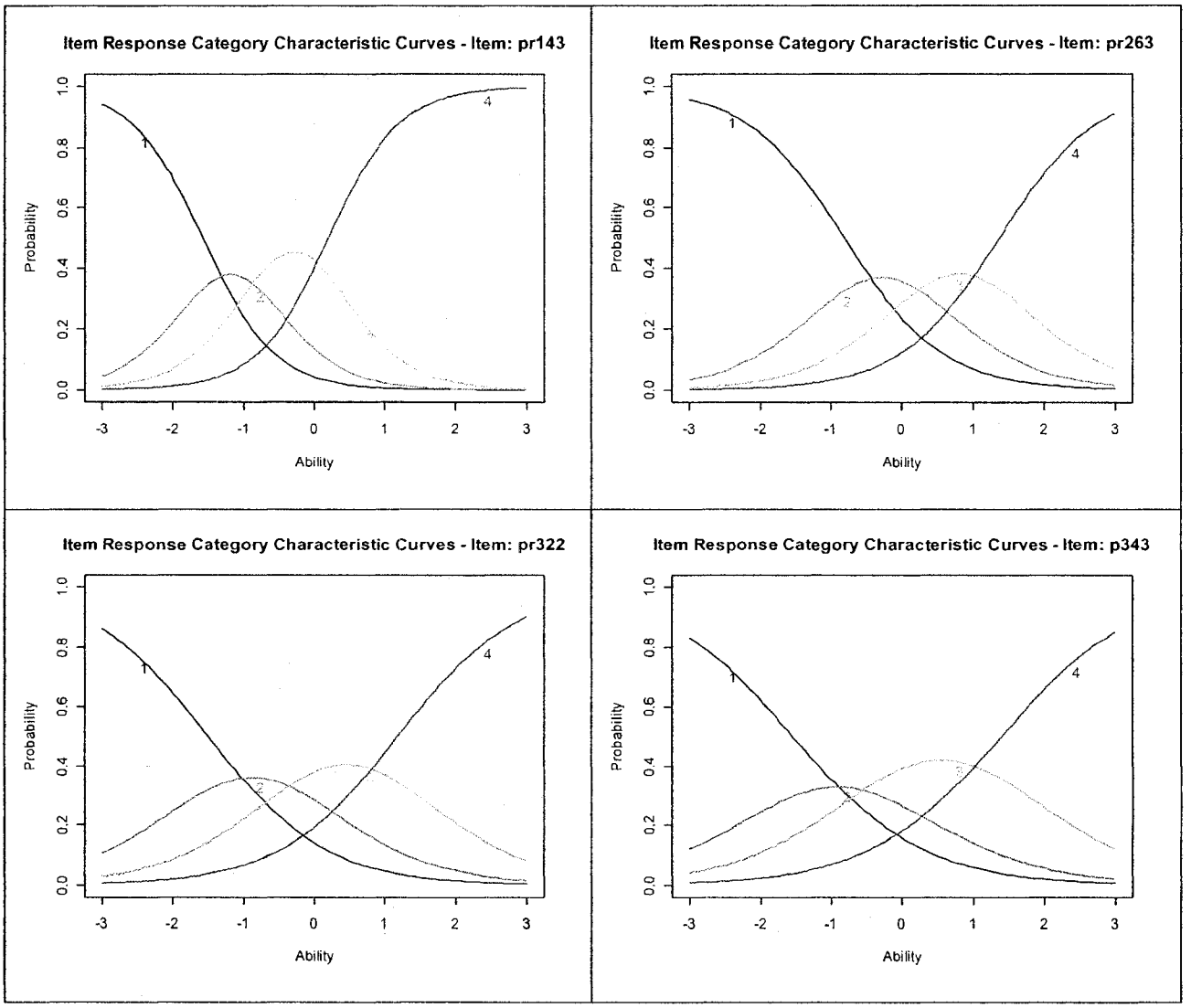

Figure 52 cont'd. Item Response Category Characteristic Curves (CCC) for COM Items (13-16)

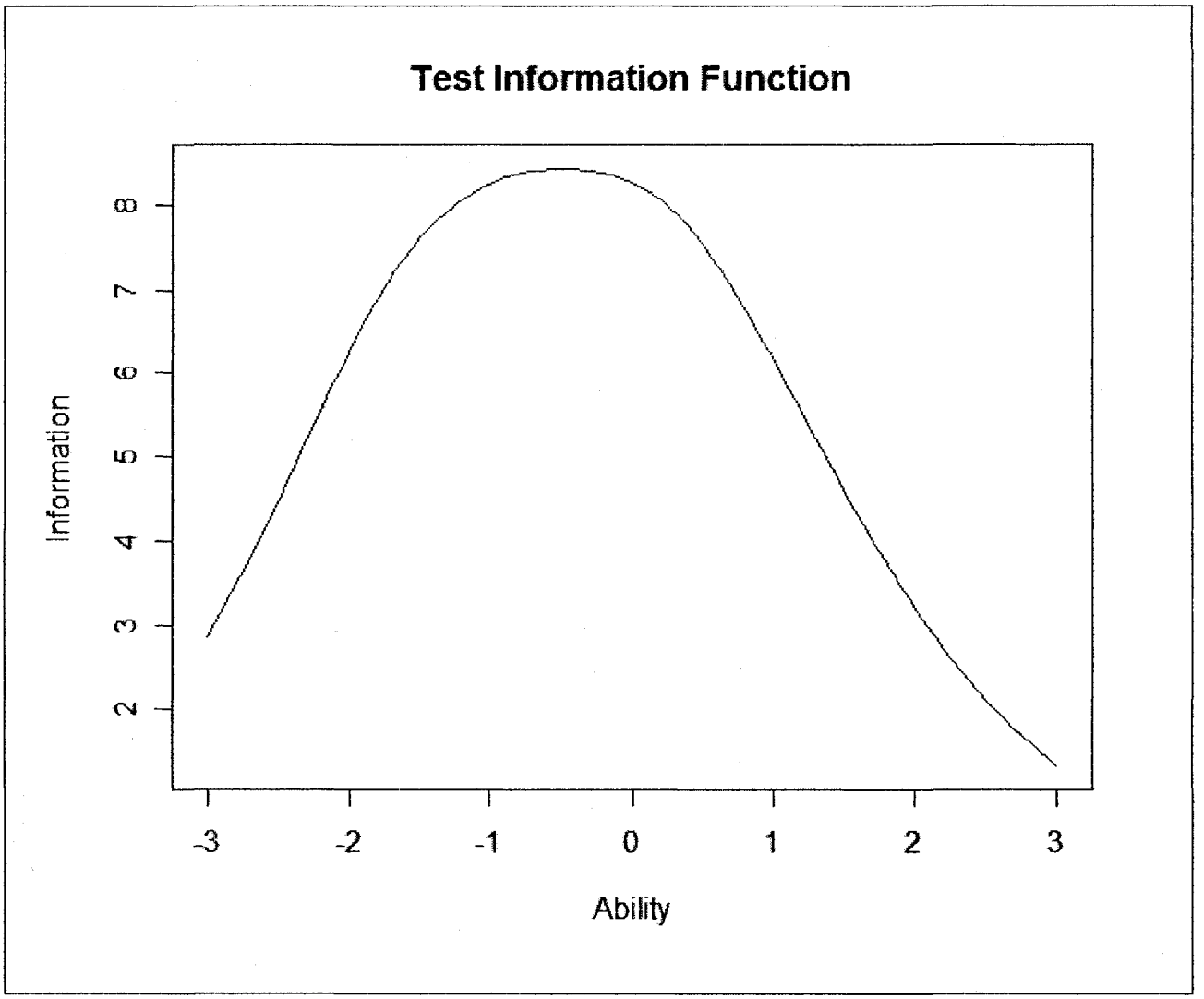

Figure 53. Test Information Function for the COM scale.

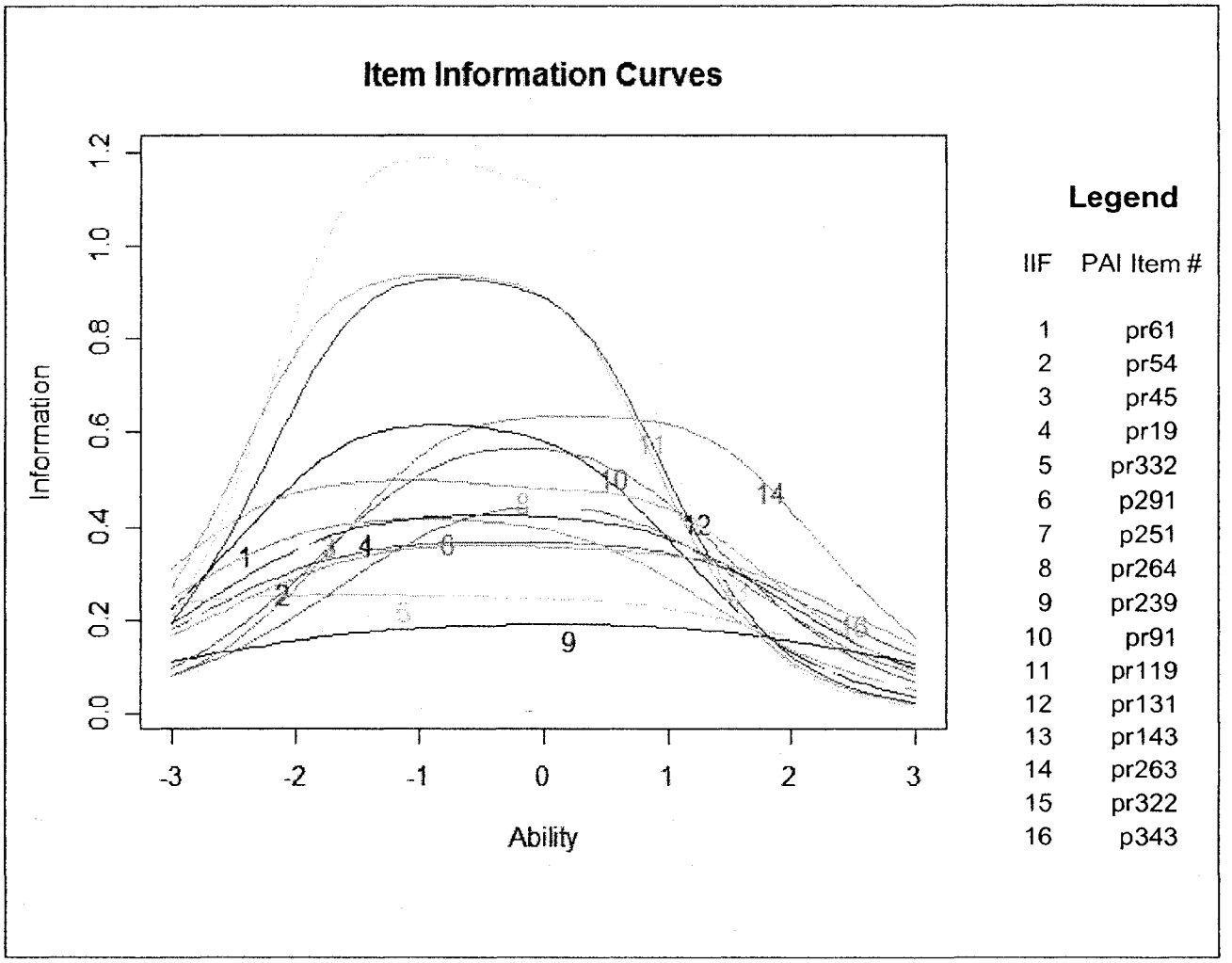

Figure 54. Item Information Functions for the COM scale. Note: p = original PAI item number, scored as is; pr = original PAI item number, reverse coded; and IIF = number corresponding to the respective, individual Item Information Function.

Table 21.1

COM: Estimated Item Parameters for the Graded Response Model

PAI Item	β1	β2	β3	α	H_g *(H)*
pr61	-1.66	-0.86	0.21	1.41	.36
pr54	-1.65	-0.28	1.18	1.09	.31
pr45	-1.99	-0.97	0.27	1.16	.32
pr19	-1.65	-0.40	0.88	1.17	.32
pr332	-3.00	-1.20	0.86	0.92	.34
p291	-0.71	0.14	0.74	1.20	.33
p251	-0.62	-0.16	0.18	0.96	.29
pr264	-2.12	-0.76	0.86	1.29	.36
pr239	-1.50	0.26	1.44	0.79	.25
pr91	-0.94	-0.18	0.62	1.34	.35
pr119	-1.73	-0.80	0.17	1.76	.38
pr131	-1.52	-0.71	0.23	1.74	.39
pr143	-1.58	-0.77	0.21	1.98	.41
pr263	-0.81	0.27	1.37	1.44	.38
pr322	-1.49	-0.24	1.19	1.21	.34
p343	-1.54	-0.27	1.39	1.08	.33
Mean	-1.53	-0.43	0.74	1.28	*(.34)*

Note. α = item slope or discrimination parameter; β = between category threshold parameter (difficulty estimate).

Table 21.2

COM: Test Information as a Function of Trait Level (Theta)

Trait Range[1]	Percent of Total Information
-3 to +3	89.22
-2 to +2	72.06
-3 to 0	52.41
0 to +3	36.81
-3 to -2	11.56
-2 to -1	19.27
-1 to 0	21.59
0 to +1	19.21
+1 to +2	11.99
+2 to +3	5.60

Note. [1] = COM trait range in *SD* units, $M = 0$, $SD = 1$; % = percent of total information or total area under the Test Information Function.

Table 21.3

COM: Item Information as a Function of Trait Level (Theta)

PAI Item	Percent of Total Information
pr61	6.60
pr54	5.52
pr45	5.44
pr19	5.80
pr332	4.90
p291	4.82
p251	3.12
pr264	7.11
pr239	3.53
pr91	5.80
pr119	9.15
pr131	8.71
pr143	10.54
pr263	7.29
pr322	6.19
p343	5.52

Note. p = original PAI item, scored as is; pr = original PAI item, reverse coded; % = percent of total information or total area under the Item Information Function.

Appendix N

PAI PD and MCMI-III correlations

Table 22

Correlations Between the New PAI PD Subscales and the MCMI-III PD Scales

		New PAI PD Scales													
		PAR	SZD	SZT	ANT[a] orig.	ANT orig.r	ANT new	BOR[b] orig.	BOR orig.r	BOR new	HIS	NAR	AVD	DEP	COM
New PAI PD Scales	PAR														
	SZD	.47													
	SZT	.67	.60												
	ANTo	.32	.03	.22											
	ANTor	.34	.03	.25	.91										
	ANTnew	.53	.23	.49	.76	.80									
	BORo	.73	.48	.76	.42	.44	.70								
	BORor	.66	.47	.79	.40	.43	.70	.95							
	BORnew	.66	.49	.79	.34	.38	.68	.90	.92						
	HIS	-.56	-.87	-.76	-.11	-.12	-.35	-.62	-.63	-.65					
	NAR	-.28	-.60	-.51	.09	.06	-.09	-.42	-.42	-.44	.61				
	AVD	.57	.61	.79	.11	.15	.39	.71	.72	.71	-.69	-.54			
	DEP	.60	.68	.82	.08	.13	.33	.68	.68	.68	-.74	-.53	.84		
	COM	-.51	-.18	-.48	-.76	-.79	-.84	-.73	-.70	-.66	.33	.13	-.38	-.35	
MCMI-III PD Scales	par	.66	.40	.54	.22	.22	.38	.54	.51	.52	-.46	-.24	.50	.47	-.36
	szd	.34	.58	.44	.08	.05	.19	.32	.33	.35	-.54	-.33	.41	.45	-.16
	szt	.55	.48	.62	.23	.26	.41	.53	.54	.55	-.54	-.28	.53	.56	-.40
	ant	.35	.10	.28	.60	.57	.60	.46	.43	.40	-.19	-.06	.23	.18	-.61
	bor	.56	.49	.69	.33	.35	.55	.72	.73	.74	-.61	-.41	.61	.59	-.56
	his	-.11	-.64	-.25	.22	.23	.10	-.08	-.10	-.13	.53	.40	-.32	-.34	-.14
	nar	.34	.06	.20	.33	.30	.36	.26	.26	.25	-.14	.25	.13	.12	-.33
	avd	.48	.64	.63	.05	.08	.24	.48	.49	.50	-.62	-.50	.63	.65	-.23
	dep	.34	.29	.53	.02	.06	.14	.41	.40	.38	-.34	-.40	.53	.58	-.20
	com	-.04	-.04	-.03	-.34	-.32	-.27	-.22	-.19	-.14	.06	.14	.00	.02	.35

Note. [a]ANT orig. = original PAI ANT scale; ANT orig.r. = original PAI ANT scale with low information items removed per IRT results; ANT new = newly created ANT scale. [b]BOR orig. = original PAI BOR scale; BOR orig.r. = original PAI BOR scale with low information items removed per IRT results; BOR new = newly created BOR scale.